Chunfeng Lian · Xiaohuan Cao ·
Islem Rekik · Xuanang Xu · Zhiming Cui (Eds.)

LNCS 13583

Machine Learning in Medical Imaging

13th International Workshop, MLMI 2022
Held in Conjunction with MICCAI 2022
Singapore, September 18, 2022, Proceedings

Springer

MOREMEDIA

Lecture Notes in Computer Science 13583

Founding Editors

Gerhard Goos
Karlsruhe Institute of Technology, Karlsruhe, Germany

Juris Hartmanis
Cornell University, Ithaca, NY, USA

Editorial Board Members

Elisa Bertino
Purdue University, West Lafayette, IN, USA

Wen Gao
Peking University, Beijing, China

Bernhard Steffen
TU Dortmund University, Dortmund, Germany

Moti Yung
Columbia University, New York, NY, USA

More information about this series at https://link.springer.com/bookseries/558

Chunfeng Lian · Xiaohuan Cao · Islem Rekik ·
Xuanang Xu · Zhiming Cui (Eds.)

Machine Learning in Medical Imaging

13th International Workshop, MLMI 2022
Held in Conjunction with MICCAI 2022
Singapore, September 18, 2022
Proceedings

Springer

Editors
Chunfeng Lian
Xi'an Jiaotong University
Xi'an, China

Islem Rekik
Istanbul Technical University
Istanbul, Turkey

Zhiming Cui
ShanghaiTech University
Pudong, China

Xiaohuan Cao
Shanghai United Imaging Intelligence Co., Ltd.
Shanghai, China

Xuanang Xu
Rensselaer Polytechnic Institute
Troy, NY, USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-21013-6 ISBN 978-3-031-21014-3 (eBook)
https://doi.org/10.1007/978-3-031-21014-3

© Springer Nature Switzerland AG 2022
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The 13th International Workshop on Machine Learning in Medical Imaging (MLMI 2022) was held in Singapore, on September 18, 2022, in conjunction with the 25th International Conference on Medical Image Computing and Computer_Assisted Intervention (MICCAI 2022).

Aiming to accelerate the intersectional research within the field of machine learning in medical imaging, the MLMI workshop provides a forum for researchers and clinicians to communicate and exchange ideas. Artificial intelligence (AI) and machine learning (ML) have significantly changed the research landscape of academia and industry. AI/ML now plays a crucial role in the medical imaging field, including, but not limited to, computer-aided detection and diagnosis, image segmentation, image registration, image fusion, image-guided intervention, image annotation, image retrieval, image reconstruction, etc. MLMI 2022 focused on major trends and challenges in this area and presented original works aiming to identify new cutting-edge techniques and their uses in medical imaging. The workshop facilitated translating medical imaging research from bench to bedside. Topics of interests included deep learning, generative adversarial learning, ensemble learning, sparse learning, multi-task learning, multi-view learning, manifold learning, and reinforcement learning, along with their applications to medical image analysis, computer-aided detection and diagnosis, multi-modality fusion, image reconstruction, image retrieval, cellular image analysis, molecular imaging, digital pathology, etc.

MLMI 2022 received a large number of papers (64 in total). Following the good practices from previous years, all the submissions underwent a rigorous double-blind peer-review process, with each paper being reviewed by at least two members of the Program Committee, composed of 38 experts in the field. Based on the reviewing scores and critiques, 48 papers were accepted for presentation at the workshop and chosen to be included in this Springer LNCS volume, which resulted in an acceptance rate of 75%. It was a tough decision and many high-quality papers had to be rejected due to the page limit of this volume.

We are grateful to all Program Committee members for reviewing the submissions and giving constructive comments. We also thank all the authors for making the workshop very fruitful and successful.

September 2022

Chunfeng Lian
Xiaohuan Cao
Islem Rekik
Xuanang Xu
Zhiming Cui

Organization

Workshop Organizers

Chunfeng Lian	Xi'an Jiaotong University, China
Xiaohuan Cao	United Imaging Intelligence, China
Islem Rekik	Istanbul Teknik Universitesi, Turkey
Xuanang Xu	Rensselaer Polytechnic Institute, USA
Zhiming Cui	ShanghaiTech University, China

Steering Committee

Dinggang Shen	ShanghaiTech University, China
Pingkun Yan	Rensselaer Polytechnic Institute, USA
Kenji Suzuki	Tokyo Institute of Technology, Japan
Fei Wang	Visa Research, USA

Program Committee

Sahar Ahmad	University of North Carolina at Chapel Hill, USA
Ulas Bagci	Northwestern University, USA
Zehong Cao	Southern Medical University and United Imaging Intelligence, China
Heang-Ping Chan	University of Michigan Medical Center, USA
Liangjun Chen	University of North Carolina at Chapel Hill, USA
Liyun Chen	Shanghai Jiao Tong University, China
Hao Guan	University of North Carolina at Chapel Hill, USA
Jiashuang Huang	Nanjing University of Aeronautics and Astronautics, China
Yuankai Huo	Vanderbilt University, USA
Khoi Huynh	University of North Carolina at Chapel Hill, USA
Caiwen Jiang	ShanghaiTech University, China
Xi Jiang	University of Electronic Science and Technology of China, China
Zhicheng Jiao	Brown University, USA
Ze Jin	Tokyo Institute of Technology, Japan
Gang Li	University of North Carolina at Chapel Hill, USA
Feihong Liu	Northwest University, USA
Jiameng Liu	ShanghaiTech University, China

Mingxia Liu	University of North Carolina at Chapel Hill, USA
Qin Liu	University of North Carolina at Chapel Hill, USA
Siyuan Liu	University of North Carolina at Chapel Hill, USA
Lei Ma	University of North Carolina at Chapel Hill, USA
Janne Nappi	Massachusetts General Hospital, USA
Chuang Niu	Rensselaer Polytechnic Institute, USA
Masahiro Oda	Nagoya University, Japan
Sanghyun Park	Daegu Gyeongbuk Institute of Science and Technology, South Korea
Kilian Pohl	SRI International, USA
Hongming Shan	Fudan University, China
Haoshen Wang	Dalian University of Technology, China
Jie Wei	Northwestern Polytechnical University, China
Han Wu	ShanghaiTech University, China
Deqiang Xiao	Beijing Institute of Technology, China
Xin Yang	The Chinese University of Hong Kong, China
Linlin Yao	Shanghai Jiao Tong University, China
Shu Zhang	Northwestern Polytechnical University, China
Xiao Zhang	Northwest University, USA
Yi Zhang	Sichuan University, China
Chongyue Zhao	University of Pittsburgh, USA
Fengjun Zhao	Northwest University, USA
Yue Zhao	Chongqing University of Posts and Telecommunications, China
Sihang Zhou	National University of Defense Technology, China
Qikui Zhu	Wuhan University, China

Contents

Function MRI Representation Learning via Self-supervised Transformer for Automated Brain Disorder Analysis 1
Qianqian Wang, Lishan Qiao, and Mingxia Liu

Predicting Age-related Macular Degeneration Progression with Longitudinal Fundus Images Using Deep Learning 11
Junghwan Lee, Tingyi Wanyan, Qingyu Chen, Tiarnan D. L. Keenan, Benjamin S. Glicksberg, Emily Y. Chew, Zhiyong Lu, Fei Wang, and Yifan Peng

Region-Guided Channel-Wise Attention Network for Accelerated MRI Reconstruction 21
Jingshuai Liu, Chen Qin, and Mehrdad Yaghoobi

Student Becomes Decathlon Master in Retinal Vessel Segmentation via Dual-Teacher Multi-target Domain Adaptation 32
Linkai Peng, Li Lin, Pujin Cheng, Huaqing He, and Xiaoying Tang

Rethinking Degradation: Radiograph Super-Resolution via AID-SRGAN 43
Yongsong Huang, Qingzhong Wang, and Shinichiro Omachi

3D Segmentation with Fully Trainable Gabor Kernels and Pearson's Correlation Coefficient 53
Ken C. L. Wong and Mehdi Moradi

A More Design-Flexible Medical Transformer for Volumetric Image Segmentation 62
Xin You, Yun Gu, Junjun He, Hui Sun, and Jie Yang

Dcor-VLDet: A Vertebra Landmark Detection Network for Scoliosis Assessment with Dual Coordinate System 72
Han Zhang, Tony C. W. Mok, and Albert C. S. Chung

Plug-and-Play Shape Refinement Framework for Multi-site and Lifespan Brain Skull Stripping 81
Yunxiang Li, Ruilong Dan, Shuai Wang, Yifan Cao, Xiangde Luo, Chenghao Tan, Gangyong Jia, Huiyu Zhou, You Zhang, Yaqi Wang, and Li Wang

A Coarse-to-Fine Network for Craniopharyngioma Segmentation 91
Yijie Yu, Lei Zhang, Xin Shu, Zizhou Wang, Chaoyue Chen, and Jianguo Xu

Patch-Level Instance-Group Discrimination with Pretext-Invariant Learning for Colitis Scoring 101
Ziang Xu, Sharib Ali, Soumya Gupta, Simon Leedham, James E. East, and Jens Rittscher

AutoMO-Mixer: An Automated Multi-objective Mixer Model for Balanced, Safe and Robust Prediction in Medicine 111
Xi Chen, Jiahuan Lv, Dehua Feng, Xuanqin Mou, Ling Bai, Shu Zhang, and Zhiguo Zhou

Memory Transformers for Full Context and High-Resolution 3D Medical Segmentation 121
Loic Themyr, Clément Rambour, Nicolas Thome, Toby Collins, and Alexandre Hostettler

Whole Mammography Diagnosis via Multi-instance Supervised Discriminative Localization and Classification 131
Qingxia Wu, Hongna Tan, Yaping Wu, Pei Dong, Jifei Che, Zheren Li, Chenjin Lei, Dinggang Shen, Zhong Xue, and Meiyun Wang

Cross Task Temporal Consistency for Semi-supervised Medical Image Segmentation 140
Govind Jeevan, S. J. Pawan, and Jeny Rajan

U-Net vs Transformer: Is U-Net Outdated in Medical Image Registration? 151
Xi Jia, Joseph Bartlett, Tianyang Zhang, Wenqi Lu, Zhaowen Qiu, and Jinming Duan

UNet-eVAE: Iterative Refinement Using VAE Embodied Learning for Endoscopic Image Segmentation 161
Soumya Gupta, Sharib Ali, Ziang Xu, Binod Bhattarai, Ben Turney, and Jens Rittscher

Dynamic Linear Transformer for 3D Biomedical Image Segmentation 171
Zheyuan Zhang and Ulas Bagci

Automatic Grading of Emphysema by Combining 3D Lung Tissue Appearance and Deformation Map Using a Two-Stream Fully Convolutional Neural Network 181
Mohammadreza Negahdar

A Novel Two-Stage Multi-view Low-Rank Sparse Subspace Clustering Approach to Explore the Relationship Between Brain Function and Structure 191
Shu Zhang, Yanqing Kang, Sigang Yu, Jinru Wu, Enze Shi, Ruoyang Wang, Zhibin He, Lei Du, and Tuo Zhang

Fast Image-Level MRI Harmonization via Spectrum Analysis 201
Hao Guan, Siyuan Liu, Weili Lin, Pew-Thian Yap, and Mingxia Liu

CT2CXR: CT-based CXR Synthesis for Covid-19 Pneumonia Classification 210
Peter Ho Hin Yuen, Xiaohong Wang, Zhiping Lin, Nikki Ka Wai Chow, Jun Cheng, Cher Heng Tan, and Weimin Huang

Harmonization of Multi-site Cortical Data Across the Human Lifespan 220
Sahar Ahmad, Fang Nan, Ye Wu, Zhengwang Wu, Weili Lin, Li Wang, Gang Li, Di Wu, and Pew-Thian Yap

Head and Neck Vessel Segmentation with Connective Topology Using Affinity Graph 230
Linlin Yao, Zhong Xue, Yiqiang Zhan, Lizhou Chen, Yuntian Chen, Bin Song, Qian Wang, Feng Shi, and Dinggang Shen

Coarse Retinal Lesion Annotations Refinement via Prototypical Learning 239
Qinji Yu, Kang Dang, Ziyu Zhou, Yongwei Chen, and Xiaowei Ding

Nuclear Segmentation and Classification: On Color and Compression Generalization 249
Quoc Dang Vu, Robert Jewsbury, Simon Graham, Mostafa Jahanifar, Shan E. Ahmed Raza, Fayyaz Minhas, Abhir Bhalerao, and Nasir Rajpoot

Understanding Clinical Progression of Late-Life Depression to Alzheimer's Disease Over 5 Years with Structural MRI 259
Lintao Zhang, Minhui Yu, Lihong Wang, David C. Steffens, Rong Wu, Guy G. Potter, and Mingxia Liu

ClinicalRadioBERT: Knowledge-Infused Few Shot Learning for Clinical Notes Named Entity Recognition 269
Saed Rezayi, Haixing Dai, Zhengliang Liu, Zihao Wu, Akarsh Hebbar, Andrew H. Burns, Lin Zhao, Dajiang Zhu, Quanzheng Li, Wei Liu, Sheng Li, Tianming Liu, and Xiang Li

Graph Representation Neural Architecture Search for Optimal Spatial/Temporal Functional Brain Network Decomposition 279
Haixing Dai, Qing Li, Lin Zhao, Liming Pan, Cheng Shi, Zhengliang Liu, Zihao Wu, Lu Zhang, Shijie Zhao, Xia Wu, Tianming Liu, and Dajiang Zhu

Driving Points Prediction for Abdominal Probabilistic Registration 288
Samuel Joutard, Reuben Dorent, Sebastien Ourselin, Tom Vercauteren, and Marc Modat

CircleSnake: Instance Segmentation with Circle Representation 298
Ethan H. Nguyen, Haichun Yang, Zuhayr Asad, Ruining Deng, Agnes B. Fogo, and Yuankai Huo

Vertebrae Localization, Segmentation and Identification Using a Graph Optimization and an Anatomic Consistency Cycle 307
Di Meng, Eslam Mohammed, Edmond Boyer, and Sergi Pujades

Coronary Ostia Localization Using Residual U-Net with Heatmap Matching and 3D DSNT 318
Milosz Gajowczyk, Patryk Rygiel, Piotr Grodek, Adrian Korbecki, Michal Sobanski, Przemyslaw Podgorski, and Tomasz Konopczynski

AMLP-Conv, a 3D Axial Long-range Interaction Multilayer Perceptron for CNNs 328
Savinien Bonheur, Michael Pienn, Horst Olschewski, Horst Bischof, and Martin Urschler

Neural State-Space Modeling with Latent Causal-Effect Disentanglement 338
Maryam Toloubidokhti, Ryan Missel, Xiajun Jiang, Niels Otani, and Linwei Wang

Adaptive Unified Contrastive Learning for Imbalanced Classification 348
Cong Cong, Yixing Yang, Sidong Liu, Maurice Pagnucco, Antonio Di Ieva, Shlomo Berkovsky, and Yang Song

Prediction of HPV-Associated Genetic Diversity for Squamous Cell Carcinoma of Head and Neck Cancer Based on ^{18}F-FDG PET/CT 358
Yuqi Fang, Jorge Daniel Oldan, Weili Lin, Travis Parke Schrank, Wendell Gray Yarbrough, Natalia Isaeva, and Mingxia Liu

TransWS: Transformer-Based Weakly Supervised Histology Image Segmentation 367
Shaoteng Zhang, Jianpeng Zhang, and Yong Xia

Contextual Attention Network: Transformer Meets U-Net 377
Reza Azad, Moein Heidari, Yuli Wu, and Dorit Merhof

Intelligent Masking: Deep Q-Learning for Context Encoding in Medical Image Analysis 387
Mojtaba Bahrami, Mahsa Ghorbani, Yousef Yeganeh, and Nassir Navab

A New Lightweight Architecture and a Class Imbalance Aware Loss Function for Multi-label Classification of Intracranial Hemorrhages 397
Prabhat Lankireddy, Chitimireddy Sindhura, and Subrahmanyam Gorthi

Spherical Transformer on Cortical Surfaces 406
Jiale Cheng, Xin Zhang, Fenqiang Zhao, Zhengwang Wu, Xinrui Yuan, John H. Gilmore, Li Wang, Weili Lin, and Gang Li

Accurate Localization of Inner Ear Regions of Interests Using Deep Reinforcement Learning 416
Ana-Teodora Radutoiu, François Patou, Jan Margeta, Rasmus R. Paulsen, and Paula López Diez

Shifted Windows Transformers for Medical Image Quality Assessment 425
Caner Özer, Arda Güler, Aysel Türkvatan Cansever, Deniz Alis, Ercan Karaarslan, and İlkay Öksüz

Multi-scale Multi-structure Siamese Network (MMSNet) for Primary Open-Angle Glaucoma Prediction 436
Mingquan Lin, Lei Liu, Mae Gorden, Michael Kass, Sarah Van Tassel, Fei Wang, and Yifan Peng

HealNet - Self-supervised Acute Wound Heal-Stage Classification 446
Héctor Carrión, Mohammad Jafari, Hsin-Ya Yang, Roslyn Rivkah Isseroff, Marco Rolandi, Marcella Gomez, and Narges Norouzi

Federated Tumor Segmentation with Patch-Wise Deep Learning Model 456
Yuqiao Yang, Ze Jin, and Kenji Suzuki

Multi-scale and Focal Region Based Deep Learning Network for Fine Brain Parcellation 466
Yuyan Ge, Zhenyu Tang, Lei Ma, Caiwen Jiang, Feng Shi, Shaoyi Du, and Dinggang Shen

Author Index 477



Function MRI Representation Learning via Self-supervised Transformer for Automated Brain Disorder Analysis

Qianqian Wang[1], Lishan Qiao[1(✉)], and Mingxia Liu[2(✉)]

[1] School of Mathematics Science, Liaocheng University, Liaocheng 252000, Shandong, China
qiaolishan@lcu.edu.cn

[2] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA
mxliu1226@gmail.com

Abstract. Major depressive disorder (MDD) is a prevalent mental health disorder whose neuropathophysiology remains unclear. Resting-state functional magnetic resonance imaging (rs-fMRI) has been used to capture abnormality or dysfunction functional connectivity networks for automated MDD detection. A functional connectivity network (FCN) of each subject derived from rs-fMRI data can be modeled as a graph consisting of nodes and edges. Graph neural networks (GNNs) play an important role in learning representations of graph-structured data by gradually updating and aggregating node features for brain disorder analysis. However, using one single GNN layer focuses on local graph structure around each node and stacking multiple GNN layers usually leads to the over-smoothing problem. To this end, we propose a transformer-based functional MRI representation learning (TRL) framework to encode global spatial information of FCNs for MDD diagnosis. Experimental results on 282 MDD patients and 251 healthy control (HC) subjects demonstrate that our method outperforms several competing methods in MDD identification based on rs-fMRI data. Besides, based on our learned fully connected graphs, we can detect discriminative functional connectivities in MDD vs. HC classification, providing potential fMRI biomarkers for MDD analysis.

Keywords: Major depressive disorder · fMRI · Transformer

1 Introduction

Major depressive disorder (MDD), also called clinical depression, is one of the most prevalent and disabling mood disorders, with the average lifetime prevalence about 12 percent [1–3]. Patients with MDD often exhibit patterns of

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_1.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 1–10, 2022.
https://doi.org/10.1007/978-3-031-21014-3_1

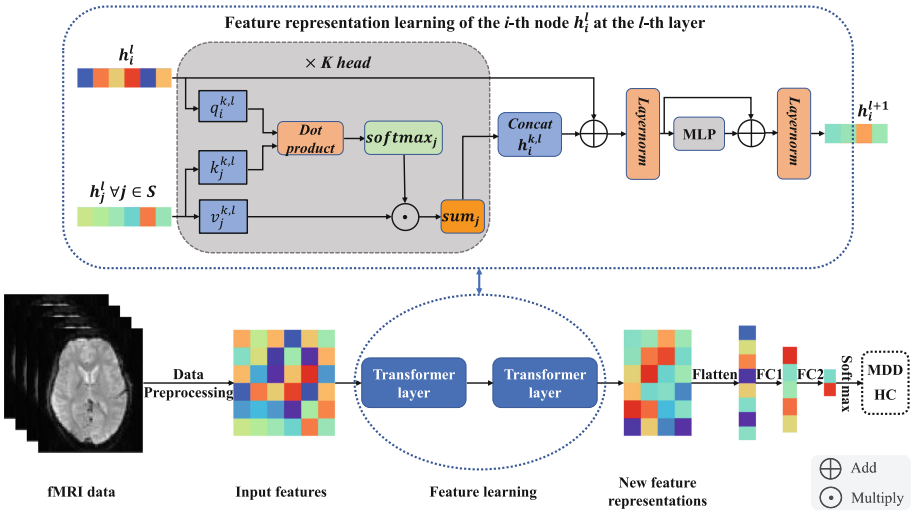


Fig. 1. Illustration of the proposed transformer-based functional MRI representation learning (TRL) framework, including three major parts: (1) preliminary graph construction, (2) transformer-based fMRI representation learning, and (3) classification.

depressed mood, loss of interest and impaired cognitive function [4,5]. Traditional diagnosis of MDD mainly generally depends on clinical manifestation and self-assessment questionnaires [6,7], which is prone to be subjective. It is highly desired to identify biomarkers for objective prognosis and diagnosis of MDD.

Resting-state functional magnetic resonance imaging (rs-fMRI), as a powerful non-invasive tool to detect brain neural activities, has been increasingly used for MDD analysis [8]. A functional connectivity network (FCN) derived from rs-fMRI data can be naturally represented as a graph consisting of nodes and edges [9–12]. Extensive studies have shown that graph neural networks (GNNs) have significant superiority in graph representation learning [13–16]. Based on rs-fMRI data, GNN can learn graph representation by gradually updating and aggregating node features, thus helping capture abnormal brain activities and discover disease-related fMRI biomarkers [9,17]. However, existing GNN-based studies usually have to face two critical challenges: (1) using one single GNN layer *cannot fully model global graph structure information*, and (2) stacking too much GNN layers usually leads to the *over-smoothing problem*.

In this work, we propose a transformer-based functional MRI representation learning framework to effectively encode global spatial information of FCNs for automated MDD detection. As shown in Fig. 1, we first extract bold-oxygen-level-dependent (BOLD) signals of rs-fMRI after data pre-processing and then construct a fully connected network/matrix for each subject. After that, the transformer layers are used to learn fMRI representation by aggregating the node features and calculating the node relationships (that is, attention scores between nodes), thus effectively capturing global topological information of brain network with the help of self-attention mechanism. Finally, two fully connected layers followed by a softmax layer are used for MDD identification. Experimental results

on 533 subjects from the REST-meta-MDD Consortium [18] validate the effectiveness of our proposed TRL method in fMRI-based MDD detection. With TRL, we can also detect some discriminative functional connectivities (FCs) related to *temporal gyrus* and *parahippocampal gyrus* that could be used as potential imaging biomarkers to facilitate MDD diagnosis in clinical practice.

2 Materials and Methodology

2.1 Subjects and Image Pre-processing

A total of 533 subjects from REST-meta-MDD Consortium [18] are used, with 282 MDD patients and 251 healthy control (HC) subject. Each rs-fMRI was acquired through a Siemens scanner, and the scanning parameters are listed as follows. repetition time (TR) $= 2,000$ ms, echo time (TE) $= 30$ ms, flip angle $= 90^o$, slice thickness $= 3.0$ mm, gap$= 1.0$ ms, time point$= 242$, voxel size $= 3.44 \times 3.44 \times 4.00\,\text{mm}^3$. Demographic information of all the studied subjects can be found in Table I of *Supplementary Materials*.

All resting-state fMRI scans were pre-processed by using the Data Processing Assistant for Resting-State fMRI (DPARSF) software [19] with a standardized pipeline [20]. We first discarded the first 10 volumes of each scan for magnetization equilibrium. Wen then performed slice timing correction, head motion correction, regression of nuisance covariates of head motion parameters, and segmentation of three tissues (*i.e.*, white matter, gray matter, and cerebrospinal fluid). Then, we normalized the fMRI data to an EPI template in the MNI space and re-sampled to have the same resolution of $3 \times 3 \times 3\,\text{mm}^3$, and performed spatial smoothing using a 6 mm full width half maximum Gaussian kernel. We finally extracted the mean rs-fMRI time series with band-pass filtering $(0.01\,Hz - 0.1\,Hz)$ of a set of 116 regions-of-interest (ROIs) defined in the automated anatomical labeling (AAL) atlas.

2.2 Proposed Method

As illustrated in Fig. 1, our framework aims to effectively encode global spatial information of brain functional connectivity networks (FCNs) by aggregating node features and calculating node relationships based on rs-fMRI, including (1) preliminary graph construction, (2) transformer-based representation learning, and (3) classification. More details are elaborated in the following.

Preliminary Graph Construction. The brain FCN of each subject can be naturally abstracted as a graph, where each node represents an anatomical ROI and each edge denotes the pairwise relationship between brain ROIs [21].

With the mean time series of each ROI, we firstly construct functional connectivity matrix. For each subject, functional connectivity is constructed by calculating Pearson correlation coefficient between mean time series of paired ROIs. Then, node feature is defined as a specific row in the functional connectivity matrix. That is, the concatenation of functional connectivity strength (*i.e.*, edge

weights) connected to a specific node will be treated as the initial node feature. Besides, in this work, we regard the FCN of each subject as a fully-connected graph and the functional connectivity strength (*e.g.*, edge weight) between ROIs will be learned via the following transformer layers.

Transformer-based fMRI Representation Learning. As shown in Fig. 1, transformer layers are used to learn new feature representation for each node/ROI in brain FCNs by modeling the dependencies between ROIs. Each transformer layer consists of (1) a multi-head self-attention module, and (2) a position-wise feed-forward network (FFN) [22]. In addition, the multi-head self-attention module and the position-wise feed-forward network (FFN) are followed by layer normalization.

Denote h_i^l and h_j^l as node representations of the i-th node and j-th node at the l-th layer, respectively. Taking feature representation learning of the i-th node as an example, h_i^l is firstly fed into multi-head self-attention module to calculate the node interrelations and also aggregate the node features. Single self-attention mechanism is mathematically formulated as follows:

$$h_i^{k,l} = Attention(h_i^l Q^{k,l}, h_j^l K^{k,l}, h_j^l V^{k,l}) = \sum\nolimits_{j \in S} w_{ij}^{k,l} h_j^l V^{k,l} \tag{1}$$

where $w_{ij}^{k,l} = softmax_j(h_i^l Q^{k,l} \cdot h_j^l K^{k,l})$ that is used to measure the attention score between the i-th node and j-th node. Also, S is the set of all nodes, and $Q^{k,l}$, $K^{k,l}$,$V^{k,l} \in R^{d_0 \times d}$ are the projection matrices of the k-th attention head at the l-th layer. Note that the self-attention mechanism is performed in a parallel manner for each node in the graph.

Then, to capture different types of spatial dependencies/relevance between nodes/ROIs, we aggregate the node representations with multiple heads. The output can be described as follows:

$$\hat{h_i^l} = LN(Concat(h_i^{1,l}, \cdots, h_i^{K,l}) W_0) \tag{2}$$

where K denotes the total number of attention heads, $W_0 \in R^{Kd \times d_0}$ is projection matrix, and each attention head is computed by the self-attention mechanism as shown in Eq. 1.

The intermediate representation obtained by multi-head self-attention module is fed into a feed-forward network (FFN), followed by residual operation and layer normalization (LN), with the updated node representation formulated as:

$$h_i^{l+1} = LN(MLP(\hat{h_i^l}) + \hat{h_i^l}) \tag{3}$$

where the Multi-Layer Perceptron (MLP) contains two linear transformations (with ReLU activation).

Classification. After obtaining the new node representations through transformer layers, we flatten the obtained embedding matrix for each subject into a subject-level vectorized representation. Then, the vectorized representation is fed into two fully connected layers, followed by a softmax layer for prediction. In the proposed method, the first and the second fully connected layers have 4,096 and 2 neurons, respectively.

Identification of Discriminative Functional Connectivity. In the transformer layers, the model can automatically learn attention scores between nodes/ROIs and produce an attention matrix at each head of each layer. Thus, we can obtain an attention matrix A for each subject by averaging the attention matrices of all heads at all layers. Regarding the brain graph of each subject as an undirected graph, we further obtain a symmetric attention matrix $\hat{A} = \frac{A+A^\top}{2}$ for each subject, and extract the upper triangle elements of the matrix to form a $6,670$-dimensional vector. Then, based on all obtained vectors, we use random forest to select the top 10 informative features and then map these features to their original functional connectivity space, thus discovering the most discriminative functional connectivities for MDD diagnosis.

Implementation Details. For the proposed TRL method, the parameters are listed as follows: K=4, d_0=116 and d=32. We optimize the TRL model using the Adam algorithm, and set the learning rate as 0.001, training epoch as 20, and mini-batch size as 32.

3 Experiment

Experimental Settings. We use a 5-fold cross-validation strategy in the experiments. Six evaluation metrics are used here, including accuracy (ACC), balanced accuracy (BAC), sensitivity (SEN), specificity (SPE), F1-Score (F1), and the area under curve (AUC).

Competing Methods. We compare our TRL method with four competing methods. (1) Clustering Coefficient (**CC**) [23]: In the CC method, we firstly generate a functional connectivity (FC) matrix for each subject using Pearson correlation coefficients. Then, clustering coefficient, which measures clustering degree of each node in the graph, is extracted as features of each FCN, and the support vector machine (SVM) with Radial Basis Function (RBF) kernel is used as the classifier. (2) **Lasso** [24]: In this method, we first extract the upper triangle elements of the constructed FC matrix for each subject to form a $6,670$-dimensional vector. Then, Lasso is used for feature selection based on vectorized features of all subjects, followed by an SVM for classification. Note that the sparsity parameter is chosen from $\{2^{-6}, \cdots, 2^6\}$ according to cross-validation. (3) Convolutional Neural Network (**CNN**) [25]: In this method, the constructed FC matrix of each subject is used as input of a CNN model. This CNN contains 3 convolutional layers and 2 fully connected layers for feature learning and classification. (4) Graph Convolutional Network (**GCN**) [26]: In this method, we first construct a KNN graph based on the FC matrix of each subject. Then, two graph convolutional layers are used to update and aggregate node representations in the graph, followed by a readout operation to generate graph-level features. The obtained representations are fed into 2 fully connected layers for feature extraction, followed by a softmax layer for classification.

Table 1. Classification results achieved by five different methods in MDD vs. HC classification. Best results are shown in bold.

Method	ACC	SEN	SPE	BAC	F1	AUC
CC	0.541	**0.674**	0.402	0.538	0.607	0.556
Lasso	0.598	0.627	0.567	0.597	0.620	0.631
CNN	0.613	0.670	0.551	0.611	0.643	0.645
GCN	0.623	0.667	0.580	0.623	0.641	0.643
TRL (Ours)	**0.636**	0.673	**0.593**	**0.633**	**0.659**	**0.674**

Experimental Results. In Table 1, we report the results of the TRL method and four competing methods in MDD vs. HC classification. From Table 1, we have several interesting observations. *On one hand*, deep learning methods (*e.g.*, CNN, GCN and TRL) are generally superior to traditional machine learning methods (*e.g.*, CC and Lasso), which demonstrates the effectiveness of deep learning methods in mining diagnosis-oriented fMRI features. *On the other hand*, our proposed TRL achieves the overall best results compared with four competing methods. For instance, compared with the method (*e.g.*, GCN) that focuses on local graph structure information around each node, the proposed TRL method achieves better performance in most cases, obtaining the ACC of 63.6%, SEN of 67.3%, SPE of 59.3%, BAC of 63.3%, F1 of 65.9%, and AUC of 67.4%. These results imply that encoding global spatial information of functional connectivity network helps boost the performance of MDD identification.

Visualization of Discriminative Functional Connectivities. We visualize the top 10 discriminative functional connectivities (FCs) identified by the proposed TRL method in Fig. 2. Note that the thickness of a line represents discriminative ability of the corresponding FC. Besides, we further list the index of the top 10 discriminative FCs and corresponding ROI names in MDD vs. HC classification in Table SII of the *Supplementary Materials*. It can be seen from Fig. 2 that several brain regions, *i.e.*, right middle temporal gyrus on temporal pole (TPOmid.R), left middle temporal gyrus on temporal pole (TPOmid.L), left inferior temporal gyrus (ITG.L), left parahippocampal gyrus (PHG.L), left superior temporal gyrus (STG.L), are identified, and right inferior temporal gyrus (ITG.R). These brain regions have been previously reported to be highly related to MDD pathology [27–29], validating the effectiveness of our method in discovering potential biomarkers for MDD identification.

4 Discussion

Influence of Three Hyperparameters. To investigate the influence of different hyperparameters, we tune three hyperparameters in the proposed TRL, including the number of attention heads (called TRL_head), the number of transformer layers (called TRL_layer) and the dimensions of transformation in Eq. (1)

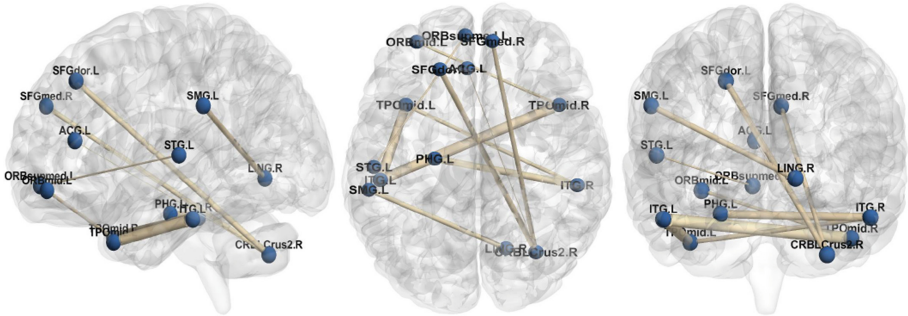


Fig. 2. Visualization of the top 10 most discriminative functional connectivities detected by our method in MDD vs. HC classification.

Table 2. Influence of three types of hyperparameters on experimental results in MDD detection, with best results shown in bold.

Hyperparameter	ACC	SEN	SPE	BAC	F1	AUC
TRL_head1	0.632	0.671	0.588	0.629	0.655	0.676
TRL_head2	0.636	**0.673**	0.593	0.633	0.659	0.674
TRL_head3	0.637	0.672	0.596	0.634	0.659	0.673
TRL_head4	**0.638**	**0.673**	0.599	**0.636**	**0.661**	**0.678**
TRL_head5	0.634	0.662	**0.603**	0.632	0.654	0.674
TRL_layer1	0.633	0.668	0.591	0.630	0.656	**0.676**
TRL_layer2	**0.636**	**0.673**	0.593	**0.633**	**0.659**	0.674
TRL_layer3	0.635	0.667	**0.599**	**0.633**	0.633	**0.676**
TRL_layer4	0.616	0.661	0.554	0.608	0.627	0.658
TRL_layer5	0.563	0.817	0.263	0.540	0.640	0.571
TRL_dim16	0.632	0.666	0.594	0.630	0.655	**0.677**
TRL_dim32	**0.636**	**0.673**	0.593	**0.633**	**0.659**	0.674
TRL_dim48	0.635	0.670	**0.596**	**0.633**	0.658	0.673
TRL_dim64	0.628	0.668	0.581	0.625	0.652	0.669
TRL_dim90	0.610	0.644	0.572	0.608	0.627	0.654

(called TRL_dim), with the corresponding results reported in Table 2. From Table 2, one can see that the TRL model using multiple attention heads generally outperforms its counterparts that use single attention head (*e.g.*, TRL_head1). This may be due to that multiple attention heads can effectively capture richer topological information of brain functional network, thus boosting the classification performance. *Besides*, the TRL achieves worse performance as the number of transformer layers increases, which implies that using too many transformer layers in TRL (*e.g.*, TRL_layer5) will not necessarily help boost the classification performance. *In addition*, as shown in Table 2, the TRL model (*e.g.*, TRL_dim16,

TRL_dim32 and TRL_dim48) achieves better performance when the dimension of transformation in Eq. 1 is relatively lower. This suggests that selecting the proper dimension of transformation in neural networks can help learn more expressive representation for classification.

Influence of Atlases for Brain Parcellation. In the main experiments, we use the AAL atlas to partition brain regions for each subject. To investigate the influence of brain parcellation on the results, we also use other brain atlases, such as Harvard-Oxford (HO) atlas, Craddock clustering 200 (CC200) atlas and Dosenbach 160 (DOS160) atlas, for ROI partition. As shown in Fig. 3, the proposed TRL method achieves comparable results using different atlases for brain parcellation, which implies that our TRL is insensitive to different brain parcellation. And the TRL method using the CC200 atlas achieves superior performance in most cases, which may be due to that fine-grained brain parcellation can capture more useful details to identify MDD subjects. These results imply that using different brain atlases for ROI partition may help capture topological features of brain network at different scales, which can be considered in future work.

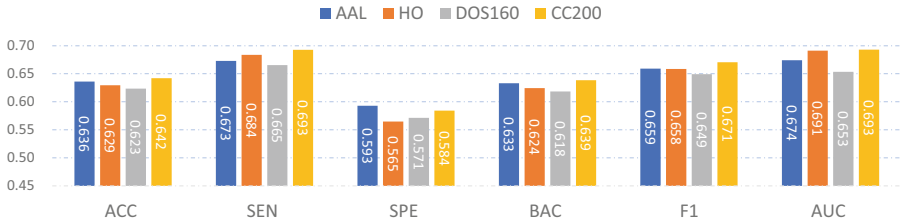


Fig. 3. Experiment results of the proposed TRL with four different brain atlases (*e.g.*, AAL, HO, DOS160 and CC200 atlases) in MDD vs. HC classification.

5 Conclusion

In this paper, we propose a transformer-based representation learning (TRL) framework for MDD identification using rs-fMRI data. We first construct a preliminary brain graph for each subject represented by rs-fMRI data, and then we employ transformer layers to encode global topological organization information of each brain graph, followed by a classification module for MDD identification. The proposed TRL can not only learn features of brain functional connectivity networks, but also helps locate those disease-related brain regions. Experimental results on a public dataset suggest the superiority of TRL over several competing methods in fMRI-based MDD identification. In current work, we only focus on modeling spatial dependencies between nodes/ROIs, ignoring modularity and hub information of graphs/FCNs. It is interesting to incorporate modularity and hub (*e.g.*, via positional encoding) information into the proposed method, which

will be our future work. Besides, using metric learning methods to further constrain the model could help learn more discriminative features, which can also be considered in the future work.

Acknowledgment. Q. Wang and L. Qiao were supported in part by Taishan Scholar Program of Shandong Province and National Natural Science Foundation of China (Nos. 62176112, 61976110 and 11931008).

References

1. Organization, W.H., et al.: Depression and other common mental disorders: global health estimates. World Health Organization, Technical report (2017)
2. Bains, N., Abdijadid, S.: Major depressive disorder. In: StatPearls [Internet]. StatPearls Publishing (2021)
3. Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R., Walters, E.E.: Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch. General Psychiatry **62**(6), 593–602 (2005)
4. Otte, C., et al.: Major depressive disorder. Nat. Rev. Dis. Primers **2**(1), 1–20 (2016)
5. Alexopoulos, G.S.: Depression in the elderly. The Lancet **365**(9475), 1961–1970 (2005)
6. Edition, F., et al.: Diagnostic and statistical manual of mental disorders. Am. Psychiatr. Assoc. **21**, 591–643 (2013)
7. Papakostas, G.I.: Managing partial response or nonresponse: switching, augmentation, and combination strategies for major depressive disorder. J. Clin. Psychiatry **70**(suppl 6), 11183 (2009)
8. Bürger, C., et al.: Differential abnormal pattern of anterior cingulate gyrus activation in unipolar and bipolar depression: an fMRI and pattern classification approach. Neuropsychopharmacology **42**(7), 1399–1408 (2017)
9. Ktena, S.I., et al.: Metric learning with spectral graph convolutions on brain connectivity networks. NeuroImage **169**, 431–442 (2018)
10. Qiao, L., Zhang, L., Chen, S., Shen, D.: Data-driven graph construction and graph learning: a review. Neurocomputing **312**, 336–351 (2018)
11. Cheng, B., Liu, M., Zhang, D., Shen, D.: Robust multi-label transfer feature learning for early diagnosis of Alzheimer's disease. Brain Imaging Behav. **13**(1), 138–153 (2019). https://doi.org/10.1007/s11682-018-9846-8
12. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. **69**, 1173–1185 (2022)
13. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.: A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst. **32**(1), 4–24 (2020)
14. Zhou, J., et al.: Graph neural networks: a review of methods and applications. AI Open **1**, 57–81 (2020)
15. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Adv. Neural Inf. Process. Syst. **30** (2017)
16. Hamilton, W.L.: Graph representation learning. Synth. Lect. Artif. Intell. Mach. Learn. **14**(3), 1–159 (2020)
17. Yao, D., et al.: A mutual multi-scale triplet graph convolutional network for classification of brain disorders using functional or structural connectivity. IEEE Trans. Med. Imaging **40**(4), 1279–1289 (2021)

18. Yan, C.C., et al.: Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. Proc. Natl. Acad. Sci. **116**(18), 9078–9083 (2019)
19. Yan, C., Zang, Y.: DPARSF: a MATLAB toolbox for "pipeline" data analysis of resting-state fMRI. Front. Syst. Neurosci. **4**, 13 (2010)
20. Yan, C.G., Wang, X.D., Zuo, X.N., Zang, Y.F.: DPABI: data processing & analysis for (resting-state) brain imaging. Neuroinformatics **14**(3), 339–351 (2016). https://doi.org/10.1007/s12021-016-9299-4
21. Sporns, O.: Graph theory methods: applications in brain networks. Dialogues Clin. Neurosci. (2022)
22. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
23. Wee, C.Y., et al.: Identification of MCI individuals using structural and functional connectivity networks. NeuroImage **59**(3), 2045–2056 (2012)
24. Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, Cham (2009). https://doi.org/10.1007/978-0-387-21606-5
25. Kawahara, J., et al.: BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage **146**, 1038–1049 (2017)
26. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
27. Kang, E.K., Lee, K.S., Lee, S.H.: Reduced cortical thickness in the temporal pole, insula, and pars triangularis in patients with panic disorder. Yonsei Med. J. **58**(5), 1018–1024 (2017)
28. Yang, Z., Guo, H., Ji, S., Li, S., Fu, Y., Guo, M., Yao, Z.: Reduced dynamics in multivariate regression-based dynamic connectivity of depressive disorder. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1197–1201. IEEE (2020)
29. Yang, X.H., et al.: Diminished caudate and superior temporal gyrus responses to effort-based decision making in patients with first-episode major depressive disorder. Progress Neuro-Psychopharmacol. Biol. Psychiatry **64**, 52–59 (2016)

Predicting Age-related Macular Degeneration Progression with Longitudinal Fundus Images Using Deep Learning

Junghwan Lee[1,6], Tingyi Wanyan[2,5,6], Qingyu Chen[3], Tiarnan D. L. Keenan[4], Benjamin S. Glicksberg[5], Emily Y. Chew[4], Zhiyong Lu[3], Fei Wang[6], and Yifan Peng[6(✉)]

[1] Columbia University, New York, USA
[2] Indiana University, Bloomington, USA
[3] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, USA
[4] National Eye Institute, National Institutes of Health, Bethesda, USA
[5] Ichan School of Medicine at Mount Sinai, New York, USA
[6] Weill Cornell Medicine, New York, USA
yip4002@med.cornell.edu

Abstract. Accurately predicting a patient's risk of progressing to late age-related macular degeneration (AMD) is difficult but crucial for personalized medicine. While existing risk prediction models for progression to late AMD are useful for triaging patients, none utilizes longitudinal color fundus photographs (CFPs) in a patient's history to estimate the risk of late AMD in a given subsequent time interval. In this work, we seek to evaluate how deep neural networks capture the sequential information in longitudinal CFPs and improve the prediction of 2-year and 5-year risk of progression to late AMD. Specifically, we proposed two deep learning models, CNN-LSTM and CNN-Transformer, which use a Long-Short Term Memory (LSTM) and a Transformer, respectively with convolutional neural networks (CNN), to capture the sequential information in longitudinal CFPs. We evaluated our models in comparison to baselines on the Age-Related Eye Disease Study, one of the largest longitudinal AMD cohorts with CFPs. The proposed models outperformed the baseline models that utilized only single-visit CFPs to predict the risk of late AMD (0.879 vs 0.868 in AUC for 2-year prediction, and 0.879 vs 0.862 for 5-year prediction). Further experiments showed that utilizing longitudinal CFPs over a longer time period was helpful for deep learning models to predict the risk of late AMD. We made the source code available at https://github.com/bionlplab/AMD_prognosis_mlmi2022 to catalyze future works that seek to develop deep learning models for late AMD prediction.

Keywords: Age-related macular degeneration · Deep learning · Convolutional neural networks · Recurrent neural networks · Transformer

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 11–20, 2022.
https://doi.org/10.1007/978-3-031-21014-3_2

1 Introduction

Age-related macular degeneration (AMD) is the leading cause of vision loss and severe vision impairment [2,16,18], which is projected to affect approximately 288 million people in the world by 2040 [21]. The annual healthcare cost incurred by AMD is about $4.6 billion in the United States imposing extreme burdens on both patients and healthcare systems.

Traditional assessments of AMD severity have been heavily dependent on the manual analysis of color fundus photographs (CFPs), which is similar to a photographic record of an ophthalmologists clinical fundus examination at the slit lamp [3]. CFPs are captured by fundus cameras and are analyzed by experts who can assess AMD severity based on multiple characteristics of a macula (e.g., presence, type, and size of drusen) [5]. The most widely used method to assess AMD severity and predict the risk of progression to late AMD is the simplified AMD severity score that was developed by the Age-Related Eye Disease Study (AREDS) Research Group [3]. The score is calculated based on the macular characteristics of CFP (or on clinical examination) from both eyes at a single time-point, and classifies an individual into 0–5 on severity scale. This severity score has been used as the current clinical standard in assessing an individuals risk of progression into late stage AMD (late AMD), based on published 5-year risks of progression to late AMD that increase from steps 0–4. However, the current clinical standard cannot incorporate longitudinal data for late AMD prediction.

While the characteristics and disease mechanisms of AMD have been well studied, there is no approved therapy for geographic atrophy that prevents slow progression of AMD to vision loss. The onset and progression of AMD can be heterogeneous between patients. Some individuals progress from early to inter mediate and to late AMD more rapidly, while others progress to late AMD more slowly [17]. Such heterogeneity requires personalized treatment planning that may be helpful in justifying medical and lifestyle interventions, vigilant home monitoring, frequent reimaging, and in planning shorter but highly powered clinical trials [15]. Meanwhile, accurately predicting late AMD risk is equally important since patient data are diverse and contain different time intervals between visits. For example, patients with early AMD could develop late AMD within a time range from a few years to several decades [13]. This requires a predictive model that focuses heavily on temporal information, which is crucial to understand disease progression [19].

To date, many machine learning methods have been used to predict late AMD onset [8], such as Logistic Regression [14], AdaBoost [1], XGBoost [23] and Random Forest [11], where the majority of these approaches use static structured features based on data from a single time-point. While these methods are straightforward and easy to implement, they typically do not capture the temporal progression information contained naturally in the data. Deep learning models have been successfully adopted in healthcare and medical tasks, and a certain amount of work has used Convolutional Neural Networks (CNNs) for addressing AMD image data [22]. Ghahramani et al. recently proposed a frame-

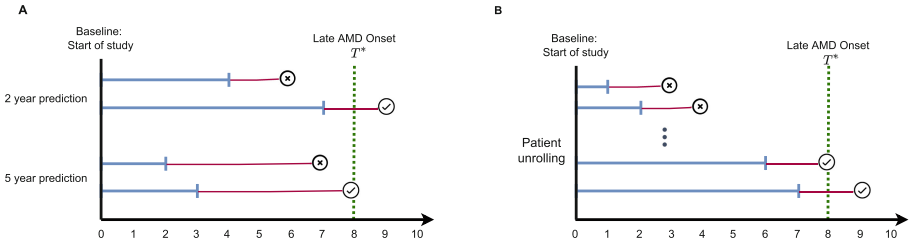


Fig. 1. Examples of 2-year and 5-year late AMD prediction task. Each CFP in individual's history is labeled 1 if late AMD onset is detected within a given prediction window (2 years and 5 years), otherwise 0. The blue lines represents the observation interval $(t_0, t_l]$, the red line represents the prediction window $(t_l, t_l + n]$. A. Prediction scenario for 2 and 5 years. B. Training scenario for a single unrolled patient. (Color figure online)

work that combines a CNN and Recurrent Neural Networks (RNN), tailored for capturing temporal progression information from CFPs for late AMD prediction [4]. However, the authors considered the data up to only three years, without the robustness of applying the model on various intervals of patient visits [13].

In this study, we used longitudinal CFPs to predict an eye as having progressed to late AMD within certain periods (2-year and 5-year). Both 2-year and 5-year predictions are common clinical scenarios. These periods were selected in advance, and it was relative to the time when the fundus photograph was taken, not to the time of the baseline visit. Specifically, for one eye, inputs are all historical CFPs and output is the probability of progression to late AMD within the specific time periods.

We proposed two deep learning models that utilize CNN with RNN and Transformer encoder respectively. We used a ResNet fine-tuned on a late AMD detection task as a fixed feature extractor, then applied a Long-Short term memory (LSTM) [10] or a Transformer encoder [20] to predict the 2-year or 5-year risk of late AMD progression. We trained and evaluated the models using longitudinal CFPs from the AREDS [7]. Models were evaluated using the area under the receiver operating characteristic curve (AUC). We compared our model with a plain ResNet, which predicts AMD progression using a single image.

Our contributions can be summarized as follows: 1) We proposed two deep learning models (CNN-LSTM and CNN-Transformer) to predict 2-year and 5-year risk of progression to late AMD; 2) We proposed a sequence unrolling strategy that can be applied to individuals having various length of sequences; 3) We showed that the proposed deep learning models outperformed baseline models that utilize only a single CFP to predict the risk of progression to late AMD; 4) We showed that longitudinal CFPs over a longer period increased the predictive performance of deep learning models to predict the risk of progression to late AMD. We also made the source code available at https://github.com/bionlplab/AMD_prognosis_mlmi2022 to catalyze future works.

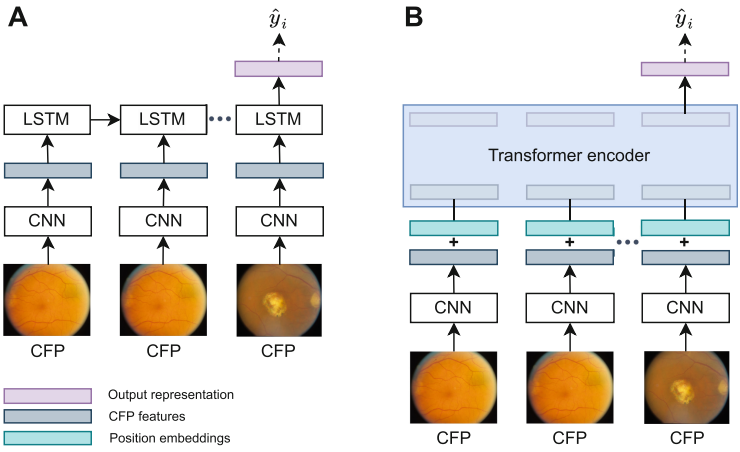


Fig. 2. Model architectures. A. CNN-LSTM. B. CNN-Transformer.

2 Methods

2.1 Definition of Late AMD Progression Prediction Task

We first formulated late AMD progression prediction task (Fig 2A). Let T^* be the 'true' time to late AMD for one participant in a study and C is right-censoring time (e.g., the end of the study). In the discrete context, we have disjoint intervals $\{t_0, t_1, ..., t_T\}$, where $T = \min(T^*, C)$ is the observed event time. Since the sequence length varies between individuals, we proposed a sequence unrolling strategy of each individual to consider all visit length (Fig 2B). Specifically, we construct $(I_0, ..., I_T)$ where $I_l = (t_0, t_l]$ is a sub-sequence of the entire observation period. In this way, a model can consider all possible length of sequences within an individual's entire history.

Given these definitions, at time t_l, our model predicts the risk of late AMD in the prediction window $(t_l, t_{l+n}]$ with longitudinal features in I_i. Here, n is the pre-selected inquiry duration. In other words, the label at time t_l is 1 if $t_{l+n} \leq T^*$; otherwise 0. In this study, we focused on 2-year ($n = 2$) and 5-year ($n = 5$) prediction because they are two common clinical scenarios [15].

This task is similar to the conditional survival probability $\frac{S(t_{l+n})}{S(t_l)}$, meaning that a participant will survive an additional n years given a survival history of t_l years. $S(t) = Pr(z > t)$ is the survival probability and z is the time for the event of late AMD.

2.2 CNN for Feature Extraction

To extract image features from CFPs, we trained a CNN model on the late AMD detection task: a task to classify a given CFP into binary label (late AMD or non-late AMD). Then, we treated this CNN as a fixed feature extractor. The CNN

was only used for feature extraction, not trained with other layers in CNN-LSTM and CNN-Transformer. There are a number of existing CNN architectures. In this study, we used ResNet [9] since it outperforms other CNN architectures in predicting AMD severity score [6]. The image features have the size of 2,048.

2.3 CNN-LSTM

In the CNN-LSTM model (Fig 1A), we used the pre-trained ResNet (Sect. 2.2) to extract features $\{f_0, ..., f_l\}$ from an individual's longitudinal CFPs in the observation window. A single fully connected layer was used to reduce dimensionality of the extracted features to the size of 256. The features were then fed into a single layer LSTM. Finally, we utilized the last output representation of LSTM for prediction.

2.4 CNN-Transformer

CNN-Transformer uses multi-layer bidirectional Transformer encoder based on the original implementation described in Vaswani et al. [20] (Fig 1B). Transformer encoder has the advantage of processing the sequence as a whole instead of processing recursively. We compute one Transformer encoder layer as:

$$\begin{aligned} \text{TransformerEncoder}(X) &= \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \\ \text{head}_i(X) &= \text{Attention}(XW_i^Q, XW_i^K, XW_i^V) \\ \text{Attention}(Q, K, V) &= softmax(\frac{QK^T}{\sqrt{d_k}})V \end{aligned}$$

where W_i^Q, W_i^K, and W_i^V are the trainable parameters and d_k is the dimension of K. We denote the input as $X = F \oplus P$, where F is the concatenation of all CFP features in the observation window of an individual and P is positional encoding. We set the number of Transformer encoder layers to 2 and the number of heads h to 8. We used the same sinusoidal function to generate positional encoding as used in [20].

2.5 Baselines

CNN-Single. In the first baseline, we used two fully connected layers on top of the extracted features by using the pre-trained ResNet: the first layer is size 256 with ReLU activation and the last layer is size 1 with sigmoid activation for making prediction. CNN-Single takes the last feature of input sequence to predict 2-year and 5-year risk prediction. We refer this baseline model to CNN-Single since it only utilizes single CFP (the last CFP in input sequence).

We also reported the prediction performance of the model proposed by Yan et al. [22], where the authors used the Inception-v3 CNN architecture to extract image features from individuals' latest visit's CFP to predict whether the eye progression time to late AMD exceeded the specific time interval.

Table 1. Characteristics of the entire cohort, 2-year prediction dataset and 5-year prediction dataset. Median value (25 percentile, 75 percentile) was reported for length of observation and CFPs per eye.

	Entire cohort	2-year	5-year
Individuals	4,315	3,477	2,876
Eyes	8,630	7,661	6,258
Eyes developed late AMD	1,768	844	428
Length of observation (year)	8(6,11)	9(6,11)	10(7,11)
CFPs	65,480	49,361	46,558
CFPs labeled as late AMD	8,422	1,240	1,573
CFPs per eye	8(5,10)	7(3,9)	8(5,10)

3 Experiment

3.1 Dataset

We used the data from AREDS, which was sponsored by the National Eye Institute of the National Institutes of Health. The data is publicly available upon request.[1] It was a 12-year multi-center prospective cohort study of the clinical course, prognosis, and risk factors of AMD, as well as a phase III randomized clinical trial to assess the effects of nutritional supplements on AMD progression. The cohort includes 4,757 participants aged 55 to 80 years, who were recruited between 1992 and 1998 at 11 retinal specialty clinics in the United States. The inclusion criteria were wide, from no AMD in either eye to late AMD in one eye. All CFPs in the cohort were labeled with 0–12 scale severity score calculated by the reading center. Individuals having missing values in the severity score or no CFP were excluded, which resulted in 4,315 individuals.

For 2-year and 5-year late AMD risk prediction, we constructed two different datasets. We first removed recurring late AMD labels from all individuals in the cohort after the first late AMD onset was detected. This is due to the irreversible nature of AMD. An individual who developed late AMD will necessarily have recurring late AMD labels after the onset of late AMD, which could make a model biased and inflate prediction performance. The individuals having observation period less than 2 and 5 years were excluded from the 2-year and 5-year prediction dataset. Then, the 2-year prediction dataset was constructed by labeling each CFP to indicate whether late AMD onset was detected within 2-year prediction window. If there was no CFP within 2-year prediction window for labeling, we excluded the CFP. The 5-year prediction dataset was constructed in the same way except that 5-year prediction window was used. The characteristics of the entire cohort and two datasets are shown in Table 1.

[1] https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1.

Table 2. AUC in predicting 2-year and 5-year risk of late AMD of all models. Data were reported as: average AUC based on 5-fold cross validation (standard deviation).

	2-year prediction	5-year prediction
Yan et al. [22]	0.810 (0.000)	0.790 (0.000)
CNN-Single	0.868 (0.012)	0.862 (0.023)
CNN-LSTM	**0.883** (0.017)	**0.879** (0.020)
CNN-Transformer	**0.879** (0.013)	**0.873** (0.020)

3.2 Experiment Setup

ResNet-101 was pre-trained to extract features from the CFPs and then the extracted features were used as input for CNN-LSTM and CNN-Transformer for late AMD risk prediction. All CFPs were resized to 256×256 and then center cropped to 224×224. Training CFPs were randomly cropped, blurred, rotated, sheared and horizontally flipped for data augmentation. ResNet-101 was trained for 30 epochs with learning rate of 0.0005 and batch size of 32. We observed that using deeper ResNet architecture than ResNet-101 did not improve the performance. CNN-LSTM and CNN-Transformer were optimized using Adam [12] with learning rate of 0.0002, batch size of 32, and epoch of 30. L2 regularization was applied to the last full-connected layer in all models to prevent overfitting. We also applied weights to the loss and the stratified mini-batch to mitigate label imbalance. All models were implemented by Tensorflow. The experiments were performed on a machine equipped with two Intel Xeon Silver 4110 CPUs and one NVIDIA RTX 2080 GPU.

We used 5-fold cross validation for evaluation and reported area under the receiver operating characteristic curve (AUC). All datasets were partitioned into training, validation, and test sets with a 3:1:1 ratio, at the participant level. This ensures that no participant was in more than one partition to avoid cross contamination between the training and test datasets. Since we observed severe imbalance between labels in the dataset (Sect. 3.1 and Table 1), which may cause instability during training, we stratified each batch during training maintaining the ratio of late AMD and non-late AMD label.

3.3 Results

Overall Prediction Performance. Table 2 shows the overall 2-year and 5-year prediction performance of all models. CNN-LSTM achieved 0.883 and 0.879 AUC in predicting 2-year and 5-year risk of late AMD. CNN-Transformer achieved 0.879 and 0.873 AUC in predicting 2-year and 5-year risk of late AMD. Both CNN-LSTM and CNN-Transformer model outperformed the baseline model in predicting 2-year and 5-year risk of late AMD. This indicates that utilizing longitudinal CFPs is helpful for the risk prediction of late AMD.

Prediction Performance Based on the Number of CFPs. We evaluated models on the subsets of datasets that only include specific number of longitudinal CFPs to investigate if the number of longitudinal CFPs affects the prediction performance. We first selected individuals having at least 5 longitudinal CFPs from test set and then sliced the longitudinal CFPs of the individuals from the last visit to build subsets of having specific number of CFPs from 2 to 5. For example, length-2 subset only includes two longitudinal CFPs sliced from the last visit. Details of the subsets are described in supplementary material.

Table 3 shows the 2-year and 5-year predictive performance based on the number of CFPs. CNN-LSTM and CNN-Transformer all showed increasing prediction performance with more longitudinal CFPs. For CNN-LSTM, AUC improved from 0.867 with 2 longitudinal CFPs to 0.873 with 5 longitudinal CFPs in both 2-year prediction and 5-year prediction. For CNN-Transformer, AUC improved from 0.862 and 0.861 with 2 longitudinal CFPs to 0.866 and 0.868 with 5 longitudinal CFPs in 2-year and 5-year prediction, respectively. This indicates that more longitudinal CFPs over longer observation period is beneficial for the risk prediction of late AMD.

Table 3. AUC in predicting 2-year and 5-year risk of late AMD of all models based on the number of CFPs. Data were reported as: average AUC based on 5-fold cross validation (standard deviation).

# of	2-year prediction		5-year prediction	
CFPs	CNN-LSTM	CNN-Transformer	CNN-LSTM	CNN-Transformer
2	0.867 (0.028)	0.862 (0.031)	0.867 (0.028)	0.861 (0.290)
3	0.870 (0.029)	0.862 (0.035)	0.870 (0.031)	0.863 (0.032)
4	0.872 (0.029)	0.864 (0.037)	0.872 (0.031)	0.866 (0.028)
5	0.873 (0.029)	0.866 (0.037)	0.873 (0.030)	0.868 (0.029)

4 Conclusion

In this work, we proposed deep learning models that utilize longitudinal color fundus photographs in individuals' histories to predict 2-year and 5-year risk of progression to late AMD. The two proposed models, CNN-LSTM and CNN-Transformer, used LSTM and Transformer encoder with CNN, respectively, and outperformed baseline models that can only utilize a single-visit CFP to predict the risk of progression to late AMD. The proposed models also showed increasing performance in predicting the risk of progression to late AMD with longitudinal CFPs over a longer period, indicating that deep learning models are effective in capturing sequential information from longitudinal CFPs in individuals' histories. Future works include the application of survival analysis to the loss objective in the models, development of end-to-end models, and integrating demographic and genetic information to further improve predictive performance.

Acknowledgments. This material is based upon work supported by the Intramural Research Programs of the National Library of Medicine and National Eye Institute at National Institutes of Health.

References

1. Chen, P., Pan, C.: Diabetes classification model based on boosting algorithms. BMC Bioinformatics **19**(1), 1–9 (2018). https://doi.org/10.1186/s12859-018-2090-9
2. Congdon, N., et al.: Causes and prevalence of visual impairment among adults in the United States. Arch. Ophthalmol. (Chicago, Ill.: 1960) **122**(4), 477–485 (2004)
3. Ferris, F.L., et al.: A simplified severity scale for age-related macular degeneration: AREDS report no. 18. Arch. Ophthalmol. (Chicago, Ill.: 1960) **123**(11), 1570–1574 (2005)
4. Ghahramani, G.C., et al.: Multi-task deep learning-based survival analysis on the prognosis of late AMD using the longitudinal data in AREDS. medRxiv (2021)
5. Graham, K.W., Chakravarthy, U., Hogg, R.E., Muldrew, K.A., Young, I.S., Kee, F.: Identifying features of early and late age-related macular degeneration: a comparison of multicolor versus traditional color fundus photography. Retina **38**(9), 1751–1758 (2018)
6. Grassmann, F., et al.: A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. Ophthalmology **125**(9), 1410–1420 (2018)
7. Age-Related Eye Disease Study Research Group.: The age-related eye disease study (AREDS): design implications AREDS report no. 1. Control. Clin. Trials **20**(6), 573 (1999)
8. Hao, S., et al.: Comparison of machine learning tools for the prediction of AMD based on genetic, age, and diabetes-related variables in the Chinese population. Regen. Ther. **15**, 180–186 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
11. Hu, C., Steingrimsson, J.A.: Personalized risk prediction in clinical oncology research: applications and practical issues using survival trees and random forests. J. Biopharm. Stat. **28**(2), 333–349 (2018)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Klein, R.: Overview of progress in the epidemiology of age-related macular degeneration. Ophthalmic Epidemiol. **14**(4), 184–187 (2007)
14. Lorenzoni, G., et al.: Comparison of machine learning techniques for prediction of hospitalization in heart failure patients. J. Clin. Med. **8**(9), 1298 (2019)
15. Peng, Y., et al.: Predicting risk of late age-related macular degeneration using deep learning. NPJ Digit. Med. **3**, 111 (2020). https://doi.org/10.1038/s41746-020-00317-z
16. Quartilho, A., Simkiss, P., Zekite, A., Xing, W., Wormald, R., Bunce, C.: Leading causes of certifiable visual loss in England and wales during the year ending 31 March 2013. Eye **30**(4), 602–607 (2016)

17. Somasundaran, S., Constable, I.J., Mellough, C.B., Carvalho, L.S.: Retinal pigment epithelium and age-related macular degeneration: a review of major disease mechanisms. Clin. Exp. Ophthalmol. **48**(8), 1043–1056 (2020)
18. Stark, K., et al.: The German AugUR study: study protocol of a prospective study to investigate chronic diseases in the elderly. BMC Geriatrics **15**(1), 1–8 (2015). https://doi.org/10.1186/s12877-015-0122-0
19. Sun, W., Rumshisky, A., Uzuner, O.: Annotating temporal information in clinical narratives. J. Biomed. Inform. **46**, S5–S12 (2013)
20. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
21. Wong, W.L., et al.: Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. The Lancet Global Health **2**(2), e106–e116 (2014)
22. Yan, Q., et al.: Deep-learning-based prediction of late age-related macular degeneration progression. Nat. Mach. Intell. **2**(2), 141–150 (2020)
23. Yu, B., et al.: SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and extreme gradient boosting. Bioinformatics **36**(4), 1074–1081 (2020)



Region-Guided Channel-Wise Attention Network for Accelerated MRI Reconstruction

Jingshuai Liu(✉), Chen Qin, and Mehrdad Yaghoobi

IDCOM, School of Engineering, University of Edinburgh, Edinburgh, UK
{J.Liu,Chen.Qin,m.yaghoobi-vaighan}@ed.ac.uk

Abstract. Magnetic resonance imaging (MRI) has been widely used in clinical practice for medical diagnosis of diseases. However, the long acquisition time hinders its development in time-critical applications. In recent years, deep learning-based methods leverage the powerful representations of neural networks to recover high-quality MR images from undersampled measurements, which shortens the acquisition process and enables accelerated MRI scanning. Despite the achieved inspiring success, it is still challenging to provide high-fidelity reconstructions under high acceleration factors. As an important mechanism in deep neural networks, attention modules have been used to improve the reconstruction quality. Due to the computational costs, many attention modules are not suitable for applying to high-resolution features or to capture spatial information, which potentially limits the capacity of neural networks. To address this issue, we propose a novel channel-wise attention which is implemented under the guidance of implicitly learned spatial semantics. We incorporate the proposed attention module in a deep network cascade for fast MRI reconstruction. In experiments, we demonstrate that the proposed framework produces superior reconstructions with appealing local visual details, compared to other deep learning-based models, validated qualitatively and quantitatively on the FastMRI knee dataset.

Keywords: MRI reconstruction · Deep learning · Region-guided channel-wise attention

1 Introduction

Magnetic resonance imaging (MRI) provides a powerful and non-invasive tool for medical diagnosis. The acquisition process is notoriously time-consuming due to physiological and hardware constraints. Undersampling k-space data is a common practice to accelerate the process, which however inevitably causes aliasing artifacts in image domain. The ill-posed problems can be modeled as,

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \lambda R(\mathbf{x}), \tag{1}$$

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 21–31, 2022.
https://doi.org/10.1007/978-3-031-21014-3_3

where $\mathbf{A}$ denotes the encoding operation, $\mathbf{y}$ is the k-space measurement, and $R(\mathbf{x})$ is a regularization on the reconstruction $\mathbf{x}$. Compressed sensing (CS) methods assume the sparsity of signals in image domain [6] or in some transformed space [11,18,22], and solve the optimization problem using iterative model-based algorithms. Nevertheless, it is challenging to hold the sparsity assumption in real scenarios and remove the aliasing artifacts via conventional methods [28], which restrains the growth of CS methods in modern MRI.

Recently, deep neural networks have been shown to perform favorably in imaging tasks [7,15,26]. Incorporating the representations of neural networks in MRI reconstruction shows superior performance in many works [10,29]. The method in [16] removes aliasing artifacts in MR images using dual magnitude and phase networks. The method in [32] introduces a primal-dual network to solve the traditional CS-MRI problem. However, the models trained with pixel-wise losses, e.g. MSE, can fail to produce sharp and realistic local details and lead to smoothed reconstructions. Generative adversarial networks (GAN) [8] exploit the adversarial game between the generator and discriminator to implicitly model the data distribution, and have been used to enhance the image quality of MRI reconstructions [19,28,30]. However, GAN-based models can potentially produce undesired details and fail to preserve faithful diagnostic features, by lacking of k-space data consistency constraints. Concurrently, attention mechanism plays an important role in vision tasks [33], which learns to capture feature dependencies to enhance the model representation capacity. Many methods leverage attention mechanism to improve MRI reconstruction quality [30]. Nevertheless, the significantly increased computational overheads limit its implementations to high-resolution features which are more closely associated with dense predictions, e.g. reconstruction tasks. Alternatively, channel-wise attention models the inter-plays between feature channels efficiently and can potentially provide better MRI reconstructions [13,17,19], whereas the spatial information is not considered, limiting the representation capacity of attention over channels.

In this paper, we introduce a novel region-guided channel-wise attention network for fast MRI reconstruction to exploit the channel-wise attention and improve the reconstruction quality. It has been incorporated with the implicitly learned spatial semantics to increase the attention diversities and gain performance boost. To provide more accurate restoration, we leverage the k-space consistency information in a densely connected network cascade and train the model in an adversarial diet. The main contributions of our work can be summarized as follows: 1) a novel region-guided channel-wise attention network for MRI reconstruction, which introduces spatial information into channel attention mechanism; 2) deriving the region-based semantic information to guide attention over channels, which increases the attention diversity and achieves performance gains; 3) by experiments, we demonstrate that the proposed method outperforms other deep learning-based approaches qualitatively and quantitatively.

2 Methods

The proposed region-guided channel-wise attention network for MRI reconstruction endows channel-wise attention with spatial diversities to enhance the reconstruction performance. We elaborate on the details as follows.

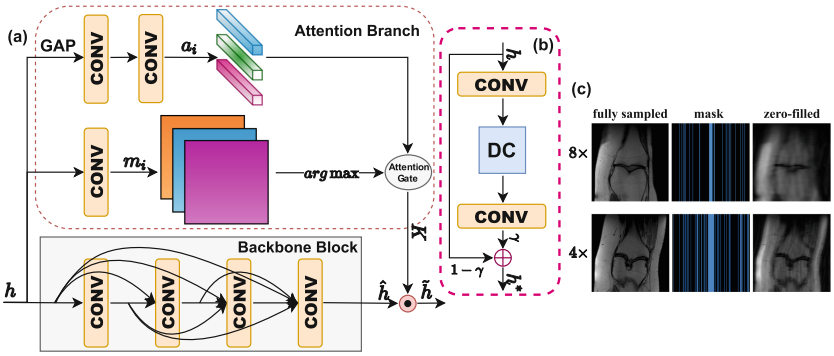


Fig. 1. Illustration of (a) RG-CAM, (b) RDCB, and (c) undersampling.

2.1 Region-Guided Channel-Wise Attention Module (RG-CAM)

A novel region-guided channel-wise attention module (RG-CAM) is proposed here to implicitly learn the spatial semantic information to guide the attention mechanism over feature channels. As displayed in Fig. 1 (a), the input features $h \in R^{C\times H\times W}$, where C and $H \times W$ denote the channel and spatial sizes, respectively attend to the backbone and attention branches. The output features $\hat{h}$ from the backbone are refined using the output K from the attention branch. The densely connected layers are used as the backbone for their effective representations [12], which can be replaced by arbitrary structures. Owing to its flexible design, RG-CAM can be easily combined with other network structures, e.g. the spatial attention modules [17,27,33], to retrieve further improvements. The details of RG-CAM are presented in the following.

Channel-Wise Attention Kernel Bank: The input h is squeezed via the global average pooling (GAP) and pass it to a non-linear mapping comprising two linear layers with a GELU activation in between, as shown in Fig. 1 (a). The output is resized to be $M \times C \times 1 \times 1$, representing M channel-wise attention kernels. Each kernel, dubbed a_i, is mapped to (0, 1) via a Sigmoid activation, where 1 means full attention to this channel and 0 denotes no attention. The kernels in bank $\{a_i\}_{i=1}^{M}$ are then incorporated together in a region-guided manner.

Spatial Guiding Mask: M spatial guiding masks $m_i \in R^{H\times W}$ are generated from h, as presented in Fig. 1 (a), to guide the implementation of channel-wise

attention. For each spatial location x, the guiding masks are normalized as below, to represent pixels from similar regions which will share an attention pattern,

$$m_j(x) = \begin{cases} 1, & j = arg\max_i m_i(x) \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Region-Guided Attention Gate: The guiding masks are amalgamated with the attention kernels to construct the region-guided attention gate $K \in R^{C\times H\times W}$, by stacking the kernels a_i via the criteria below,

$$K(x) \leftarrow a_i, \text{ if } m_i(x) = 1. \tag{3}$$

Due to the spatial invariance of convolutional networks, regions with similar semantics potentially have the same values in m_i and share the attention patterns. The final output $\tilde{h}$ of RG-CAM is given as follows,

$$\tilde{h} = K \odot \hat{h}, \tag{4}$$

where $\odot$ is the element-wise multiplication. RG-CAM endows the spatial diversity to attention over channels in a flexible, efficient, and light-weight manner. Compared to the backbone branch, the parameters and computational costs of the attention branch are negligible due to the GAP operation. It requires neither labeled data nor extra supervision, and is end-to-end trained. We use the Softmax trick [14] to enable the gradient propagation w.r.t the guiding masks and select $M = 8$ for all RG-CAMs.

2.2 Residual Data Consistency Block (RDCB)

GAN-based models have been proven to generate photo-realistic images for MRI reconstruction [4,9,28]. However, the lack of k-space constraints can lead to irrelevant details [3] and degradation in the reconstruction quality. To encourage more consistent reconstructions with the measurements $\mathbf{y}$, we propose a residual data consistency block (RDCB) to leverage the k-space information to "correct" the intermediate predictions. DC methods conventionally reduce the feature channels to handle complex-valued signals, i.e. using 2 channels, which can be detrimental to the model performance due to the bottleneck design. For example, the input and output channel sizes of the first convolution in (5) and Fig. 1 (b) are 16 and 2, representing the feature maps and complex signals, and vice versa for the second convolution. Instead, we take advantage of residual learning in the feature space to facilitate feature propagation in DC blocks. As presented in Fig. 1 (b), the resultant RDCB can be formulated as below,

$$h^* \leftarrow (1-\gamma) \times h + \gamma \times \text{conv}(DC(\text{conv}(h), \mathbf{y}; \mathbf{A})), \tag{5}$$

where h and h^* denote the input and output features, γ is a trainable parameter, and DC refers to the data consistency operation [21,25] which is given by,

$$DC(\mathbf{x}, \mathbf{y}; \mathbf{A}) = \mathbf{x} - \mathbf{A}^H(\mathbf{A}\mathbf{x} - \mathbf{y}). \tag{6}$$

In Sect. 3.3, we demonstrate that the slight modification delivers performance improvements, showing the efficacy of the proposed residual design.

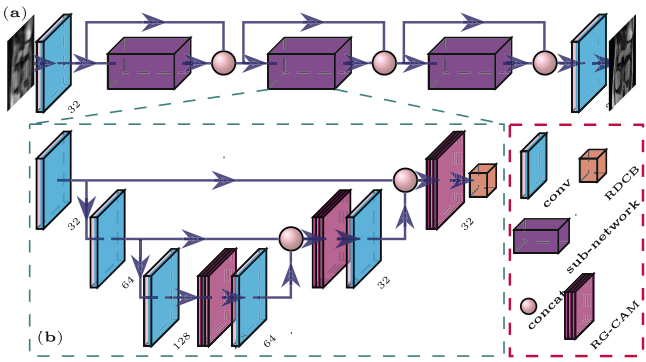


Fig. 2. Illustration of (a) densely connected reconstruction cascade and (b) U-shaped sub-network.

2.3 Densely Connected Reconstruction Cascade

Deep cascaded networks are shown to yield higher performance in MRI reconstruction [1,5,16,24], by virtue of the representation power of deep structures. Inspired by dense connections in [12], we propose a densely connected reconstruction cascade to facilitate feature reuse and transmission. As shown in Fig. 2 (a), the current predictions are collected with the outputs from the preceding sub-networks via concatenation and passed as input to the following model structure. The collection of output features from all sub-networks are fused to give the final reconstruction. The framework takes the zero-filled as input and adopts five U-shaped sub-networks as illustrated in Fig. 2 (b).

2.4 Objective Function

We adopt the L_1 metric and structural similarity index (SSIM) L^{SSIM} to measure the reconstruction errors. We also train the model with an adversarial loss L_{adv} [20] to encourage sharp details. The total objective is given as below, with practically selecting $\alpha = 0.4$ and $\lambda_{adv} = 0.01$ in our experiments,

$$L = (1-\alpha)L_1(G,s) + \alpha L^{SSIM}(G,s) + \lambda_{adv}L_{adv}(G,s), \tag{7}$$

where G and s refer to the reconstruction and reference.

3 Experiment and Results

3.1 Data and Implementation Details

We use the FastMRI single-coil knee cases [31] to conduct the experiments. Two random sampling masks are used with acceleration factors of 8× and 4×, as shown in Fig. 1 (c). We use two channels to represent complex-valued signals. The model is trained for 35 epochs with a batch size of 5, using an Adam optimizer with β_1=0.5, β_2=0.999, and a learning rate of 2×10^{-4}. The method was implemented in PyTorch on a NVIDIA RTX 3090 GPU.

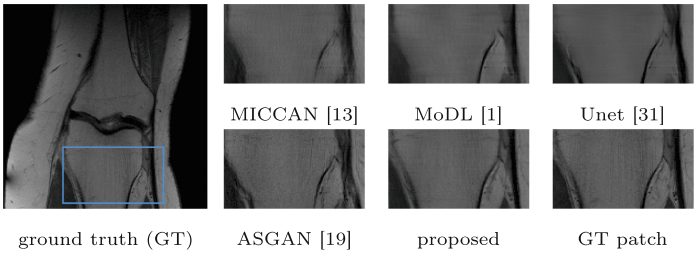


Fig. 3. Comparison results of 8× accelerated MRI reconstruction.

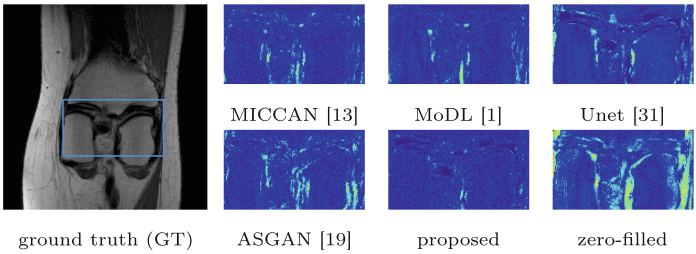


Fig. 4. Error maps (2× amplified) of 8× accelerated MRI reconstruction.

3.2 Comparison Experiments

We present the comparison results with other state-of-the-art methods: MICCAN [13], MoDL [1], FastMRI Unet [31], and ASGAN [19]. MICCAN and MoDL both adopt deep network cascades for reconstruction, where a channel-wise attention module is used in MICCAN. ASGAN adopts a GAN-based framework and performs attention selection for feature channels. We present the reconstructions using 8× and 4× acceleration factors in Fig. 3. It can be observed that the proposed method produces more faithful results with rich detailed structures, compared to other results. From the residual maps in Fig. 4, it is also shown that the proposed method restores the undersampled images with a higher accuracy, particularly at a high acceleration rate. Quantitatively, we use PSNR and SSIM as reconstruction error measurement and adopt FID and KID [2] for the visual evaluation. Table 1 lists the evaluation results, where the proposed method outperforms other competing approaches at both sampling rates (p-value$\ll$0.05). Due to the encoding-decoding structure, our method shows competitive inference speed, potentially enabling real-time reconstruction. Additionally, we replace conventional convolutions in sub-networks with separable (depth-wise+point-wise) convolutions [23] to strike a better accuracy-latency trade-off. The resultant variant "proposed-S" with fewer parameters is adopted to further verify that the superior performance is attributable to the proposed model components and structure, and not simply due to the size of model. We report the ablation results in the following section.

Table 1. Quantitative evaluation on accelerated MRI reconstruction. 4× accelerated reconstructions have the same run-time and model size as 8×.

	Method	PSNR↑	SSIM↑	FID↓	KID↓	Run-time (s)↓	Size (MB)↓
8×	Proposed	**28.65**	**0.758**	**74.26**	**0.012**	0.049	44.5
	Proposed-S	28.12	0.747	80.00	0.014	0.041	22.9
	ASGAN [19]	25.45	0.638	104.34	0.036	0.056	17.0
	Unet [31]	25.82	0.703	160.35	0.121	**0.013**	10.5
	MoDL [1]	27.13	0.620	143.65	0.080	0.091	22.3
	MICCAN [13]	26.61	0.642	180.66	0.146	0.043	**10.1**
	Zero-filled	20.54	0.388	423.32	0.533	–	–
4×	Proposed	**32.22**	**0.854**	**57.18**	**0.003**	–	–
	Proposed-S	32.01	0.850	58.15	0.004	–	–
	ASGAN [19]	27.73	0.711	82.18	0.016	–	–
	Unet [31]	28.35	0.771	118.07	0.061	–	–
	MoDL [1]	30.34	0.745	98.86	0.042	–	–
	MICCAN [13]	30.11	0.711	99.44	0.040	–	–
	Zero-filled	23.94	0.486	255.06	0.239	–	–

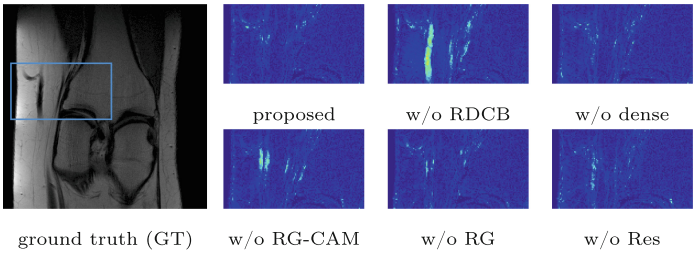


Fig. 5. Ablation residual maps (2× amplified) of 8× accelerated reconstruction.

3.3 Ablation Analysis

We conduct ablation experiments to evaluate the role of model components. We present the ablation results in Table 2, where "w/o RDCB", "w/o dense", and "w/o RG-CAM" respectively mean the proposed method without RDCB, dense network connections, and RG-CAM. For fair comparisons, we use feature repetition to maintain the channel size and model parameters, when removing the dense connections. The backbone in RG-CAM contains significantly more parameters than the attention branch, and it is maintained for a fair comparison. The convolutional layers in RDCB are also kept for the same reason. The results show that the proposed model components consistently improve the reconstruction performance in terms of evaluation metrics (p-value$\ll$0.05).

To further demonstrate the role of RG-CAM, the variant, dubbed "w/o RG", adopts a single kernel to perform channel-wise attention without using the spatial

Table 2. Ablation studies on model components at 8× acceleration factor.

Method	PSNR↑	SSIM↑	FID↓	KID↓
Proposed	**28.65**	**0.758**	**74.26**	**0.012**
w/o RDCB	27.33	0.731	83.43	0.017
w/o dense	28.03	0.748	81.25	0.015
w/o RG-CAM	28.29	0.753	78.67	0.014
w/o RG	28.26	0.754	77.40	0.013
w/o Res	28.34	0.748	80.13	0.015

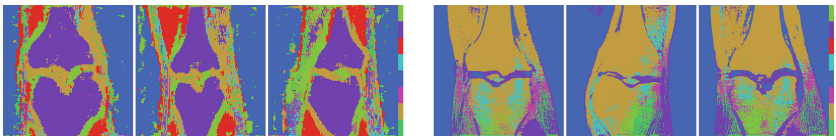


Fig. 6. Visualization of guiding masks from (left) penultimate decoding level of the 4-th sub-network and (right) last decoding level of the 5-th sub-network.

guiding masks, similar to [13,17,27]. From Table 2, we found that the incorporation of the region-guided mechanism enhances the model performance, comparing "w/o RG" and "proposed". In contrast, adopting conventional channel-wise attention fails to notably gain performance boost, comparing "w/o RG-CAM" and "w/o RG", which suggests the usefulness of the proposed region-guided method. Additionally, the variant "w/o Res" refers to the removal of the residual structure in RDCB. The results in Table 2 confirm its efficacy in delivering performance gains. From the residual maps in Fig. 5, it is shown that the proposed method introduces more accurate reconstructions, compared to other candidates.

3.4 Region-Guided Mask Visualization

To visualize the spatial information learned in RG-CAM, we conflate the guiding masks where non-zero pixels are shaded in different colors. From the heat maps in Fig. 6, we can observe clear "segmentations" of different structures. It indicates that the spatial semantics are implicitly captured by "clustering" pixels from similar regions, which share the same attention patterns. It is worth noting that the region-based guiding information is learned and incorporated in channel recalibration without requiring any annotations for supervision.

4 Conclusions

In this paper, a novel region-guided channel-wise attention network is introduced for accelerated MRI reconstruction, which adopts an efficient and lightweight structure to simultaneously make use of the channel-wise attention and

the implicitly learned spatial semantics. Incorporated with network dense connections and data consistency priors, it is demonstrated that the proposed method yields superior reconstruction performance at different acceleration factors, which can considerably shorten the MRI scanning time. For the future works, we plan to apply our method to other anatomical structures, and extend it to dynamic MRI reconstruction.

References

1. Aggarwal, H., Mani, M., Jacob, M.: MoDL: model-based deep learning architecture for inverse problems. IEEE Trans. Med. Imaging **38**(2), 394–405 (2019). https://doi.org/10.1109/TMI.2018.2865356
2. Bińkowski, M., Sutherland, D., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: International Conference on Learning Representations (2018)
3. Chen, S., Sun, S., Huang, X., Shen, D., Wang, Q., Liao, S.: Data-consistency in latent space and online update strategy to guide GAN for fast MRI reconstruction. In: Deeba, F., Johnson, P., Würfl, T., Ye, J.C. (eds.) MLMIR 2020. LNCS, vol. 12450, pp. 82–90. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61598-7_8
4. Deora, P., Vasudeva, B., Bhattacharya, S., Pradhan, P.M.: Structure preserving compressive sensing MRI reconstruction using generative adversarial networks. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2020)
5. Duan, J., et al.: VS-Net: variable splitting network for accelerated parallel MRI reconstruction. In: Medical Image Computing and Computer Assisted Intervention, vol. 11767, pp. 713–722. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_78
6. Fair, M., Gatehouse, P., DiBella, E., Firmin, D.: A review of 3D first-pass, whole-heart, myocardial perfusion cardiovascular magnetic resonance. J. Cardiovasc. Magn. Reson. (2015). https://doi.org/10.1186/s12968-015-0162-9
7. Gatys, L., Ecker, A., Bethge, M.: Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423 (2016). https://doi.org/10.1109/CVPR.2016.265
8. Goodfellow, I., et al.: Generative adversarial networks. Adv. Neural Inf. Process. Syst. **27**, 2672–2680 (2014)
9. Guo, Y., Wang, C., Zhang, H., Yang, G.: Deep attentive Wasserstein generative adversarial networks for MRI reconstruction with recurrent context-awareness. In: Medical Image Computing and Computer Assisted Intervention, vol. 12262, pp. 166–177. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_17
10. Hammernik, K., et al.: Learning a variational network for reconstruction of accelerated MRI data. Magn. Reson. Med. **79**(6), 3055–3071 (2018). https://doi.org/10.1002/mrm.26977
11. Hong, M., Yu, Y., Wang, H., Liu, F., Crozier, S.: Compressed sensing MRI with singular value decomposition-based sparsity basis. Phys. Med. Biol. **56**, 6311–6325 (2021)
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269 (2017)

13. Huang, Q., Yang, D., Wu, P., Qu, H., Yi, J., Metaxas, D.: MRI reconstruction via cascaded channel-wise attention network. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1622–1626 (2019). https://doi.org/10.1109/ISBI.2019.8759423
14. Kenji, I., Kuroki, R., Uchida, S.: Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In: International Conference on Computer Vision Workshop, ICCVW 2019, pp. 4176–4185 (2019)
15. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25** (2012)
16. Lee, D., Yoo, J., Tak, S., Ye, J.: Deep residual learning for accelerated MRI using magnitude and phase networks. IEEE Trans. Biomed. Eng **65**(9), 1985–1995 (2018)
17. Li, G., Lv, J., Wang, C.: A modified generative adversarial network using spatial and channel-wise attention for CS-MRI reconstruction. IEEE Access **9**, 83185–83198 (2021)
18. Lingala, S., Jacob, M.: Blind compressive sensing dynamic MRI. IEEE Trans. Med. Imaging **32**(6), 1132–1145 (2013)
19. Liu, J., Yaghoobi, M.: Fine-grained MRI reconstruction using attentive selection generative adversarial networks. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1155–1159 (2021)
20. Mao, X., Li, Q., Xie, H., Lau, R., Wang, Z., Smolley, S.: Least squares generative adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2813–2821 (2017)
21. Pezzotti, N., Yousefi, S., Elmahdy, M., van Gemert, J., Schülke, C., Doneva, M., et al.: An adaptive intelligence algorithm for undersampled knee MRI reconstruction. arXiv e-prints arXiv:2004.07339 (2020)
22. Ravishankar, S., Bresler, Y.: MR image reconstruction from highly undersampled k-space data by dictionary learning. IEEE Trans. Med. Imaging **30**(5), 1028–1041 (2011)
23. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.: MobileNetV2: inverted residuals and linear bottlenecks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
24. Schlemper, J., Caballero, J., Hajnal, J., Price, A., Rueckert, D.: A deep cascade of convolutional neural networks for dynamic MR image reconstruction. IEEE Trans. Med. Imaging **37**(2), 491–503 (2018)
25. Sriram, A., et al.: End-to-end variational networks for accelerated MRI reconstruction. In: Medical Image Computing and Computer Assisted Intervention - MICCAI, vol. 12262, pp. 64–73. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_7
26. Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J.: Image inpainting via generative multi-column convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 331–340 (2018)
27. Woo, S., Park, J., Lee, J., Kweon, I.: CBAM: convolutional block attention module. CoRR abs/1807.06521 arXiv:1807.06521 (2018)
28. Yang, G., et al.: DAGAN: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. IEEE Trans. Med. Imaging **37**(6), 1310–1321 (2018)
29. Yang, Y., Sun, J., Li, H., Xu, Z.: Deep ADMM-Net for compressive sensing MRI. Adv. Neural Inf. Process. Syst. **29** (2016)

30. Yuan, Z., Jiang, M., Wang, Y., Wei, B., et al.: SARA-GAN: self-attention and relative average discriminator based generative adversarial networks for fast compressed sensing MRI reconstruction. Front. Neuroinform. 1–12 (2020)
31. Zbontar, J., Knoll, F., Sriram, A., Muckley, M., Bruno, M., et al.: FastMRI: an open dataset and benchmarks for accelerated MRI. CoRR abs/1811.08839 arXiv:1811.08839 (2018)
32. Zhang, C., Liu, Y., Shang, F., Li, Y., Liu, H.: A novel learned primal-dual network for image compressive sensing. IEEE Access **9**, 26041–26050 (2021). https://doi.org/10.1109/ACCESS.2021.3057621
33. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, vol. 97, pp. 7354–7363 (2019)



Student Becomes Decathlon Master in Retinal Vessel Segmentation via Dual-Teacher Multi-target Domain Adaptation

Linkai Peng[1], Li Lin[1,2], Pujin Cheng[1], Huaqing He[1], and Xiaoying Tang[1,3(✉)]

[1] Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China
tangxy@sustech.edu.cn

[2] Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR, China

[3] Jiaxing Research Institute, Southern University of Science and Technology, Jiaxing, China
tangxy@sustech.edu.cn

Abstract. Unsupervised domain adaptation has been proposed recently to tackle the so-called domain shift between training data and test data with different distributions. However, most of them only focus on single-target domain adaptation and cannot be applied to the scenario with multiple target domains. In this paper, we propose RVms, a novel unsupervised multi-target domain adaptation approach to segment retinal vessels (RVs) from multimodal and multicenter retinal images. RVms mainly consists of a style augmentation and transfer (SAT) module and a dual-teacher knowledge distillation (DTKD) module. SAT augments and clusters images into source-similar domains and source-dissimilar domains via Bézier and Fourier transformations. DTKD utilizes the augmented and transformed data to train two teachers, one for source-similar domains and the other for source-dissimilar domains. Afterwards, knowledge distillation is performed to iteratively distill different domain knowledge from teachers to a generic student. The local relative intensity transformation is employed to characterize RVs in a domain invariant manner and promote the generalizability of teachers and student models. Moreover, we construct a new multimodal and multicenter vascular segmentation dataset from existing publicly-available datasets, which can be used to benchmark various domain adaptation and domain generalization methods. Through extensive experiments, RVms is found to be very close to the target-trained Oracle in terms of segmenting the RVs, largely outperforming other state-of-the-art methods.

Keywords: Multi-target domain adaptation · Dual teacher · Knowledge distillation · Style transfer · Retinal vessel segmentation

L. Peng and L. Lin—Contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_4.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 32–42, 2022.
https://doi.org/10.1007/978-3-031-21014-3_4

1 Introduction

Multimodal ophthalmic images can be effectively employed to extract retinal structures, identify biomarkers, and diagnose diseases. For example, fundus images can depict critical anatomical structures such as the macula, optic disc, and retinal vessels (RVs) [3,11]. Optical coherence tomography angiography (OCTA) can efficiently and accurately generate volumetric angiography images [14,20]. These two modalities can both deliver precise representations of vascular structures within the retina, making them popular for the diagnoses of eye-related diseases. These years, advance in the imaging technologies brings new ophthalmic modalities including optical coherence tomography (OCT) [10], widefield fundus [4], and photoacoustic images [9]. These new modalities also provide important information of RVs and are likely to enhance the disease diagnosis accuracy. In retinal disease analysis, RV segmentation is a very important pre-requisite. Manual delineation is the most accurate yet highly labor-intensive and time-consuming manner, especially for RV segmentation across multiple modalities. One plausible solution is to manually label RVs on images of a single modality and then transfer the labels to other modalities of interest via advanced image processing and deep learning techniques. However, researches have shown that deep learning models trained on one domain generally perform poorly when tested on another domain with different data distribution, because of domain shift [22]. This indicates that segmentation models trained on existing modalities will have low generalizability on new modalities. Furthermore, there also exist variations within images of the same modality because of different equipments and imaging settings, posing more challenges to a generalized RV segmentation model for images with cross-center or cross-modality domain shift.

To address this issue, domain adaptation has been explored, among which unsupervised domain adaptation (UDA) has gained the most favor in recent years. For instance, Javanmardi et al. [12] combined U-net [21] with a domain discriminator and introduced adversarial training based on DAAN [6]. Wang et al. [28] adopted this design to joint optic disc and cup segmentation from fundus images. These two methods focus on domain shift with small variations (i.e., cross-center domain shift). To tackle domain shift with large variations (i.e., cross-modality domain shift), Cai et al. [2] incorporated CycleGAN [32] into its own network and employed a shape consistency loss to ensure a correct translation of semantics in medical images. Dou et al. [5] designed a cross-modality UDA framework for cardiac MR and CT image segmentation by performing DA only at low-level layers, under the assumption that domain shift mainly lies in low-level characteristics. Zhang et al. [30] presented a Noise Adaptation Generative Adversarial Network for RV segmentation and utilized a style discriminator enforcing the translated images to have the same noise patterns as those in the target domain. Peng et al. [19] employed disentangled representation learning to disentangle images into content space and style space to extract domain-invariant features. These approaches are nevertheless designed to handle only a specific type of domain shift and may have limitations when applied to multimodal and multicenter ($m\&m$) ophthalmic images.

We here propose a novel unsupervised multi-target domain adaptation (MTD A) approach, RVms, to segment RVs from $m\&m$ retinal images. RVms mainly consists of two components: a style augmentation and transfer (SAT) module and a dual-teacher knowledge distillation (DTKD) module. SAT uses Bézier transformation to augment source images into diverse styles. By adjusting the interpolation rate, source-style images are gradually transferred to target styles via Fourier Transform. We then cluster images into source-similar domains D_{sim} and source-dissimilar domains D_{dis} according to the vessel-background relative intensity information. DTKD utilizes the augmented and transformed data to train two teachers, one for D_{sim} and the other for D_{dis}. Then knowledge distillation [8] is performed to iteratively distill different domain knowledge from teachers to a generic student. In addition, the local relative intensity transformation (LRIT) is employed to characterize RVs in a domain invariant manner and to improve the generalizability of teachers and student models.

The main contributions of this work are four-fold: (1) We propose a novel MTDA framework for RV segmentation from $m\&m$ retinal images. To the best of our knowledge, this is the first work that explores unsupervised MTDA for medical image segmentation across both modalities and centers. (2) MTDA features itself with style augmentation, style transfer, and dual-teacher knowledge distillation, which effectively tackles both cross-center and cross-modality domain shift. (3) Extensive comparison experiments are conducted, successfully identifying the proposed pipeline's superiority over representative state-of-the-art (SOTA) methods. (4) We construct a new $m\&m$ vascular segmentation dataset from existing publicly-available datasets, which can be used to benchmark various domain adaptation and domain generalization methods. Our code and dataset are publicly available at https://github.com/lkpengcs/RVms.

2 Methodology

The proposed RVms framework is shown in Fig. 1, which consists of a SAT module and a DTKD module.

2.1 Definition

We define the source domain as $S = \{x_i^s, y_i^s\}_{i=1}^{N^s}$, where x_i^s is the i-th input image, y_i^s is the corresponding RV segmentation label, and N^s is the total number of source domain images. The target domains are denoted as $T = \{T_1, T_2, \ldots T_n\}$, where $T_t = \left\{x_j^t, y_j^t\right\}_{j=1}^{N^t}$, x_j^t is the j-th image of the t-th target domain, y_j^t is the corresponding RV segmentation label, and N^t is the total number of images from the t-th target domain. From SAT, we obtain images in source-similar domains D_{sim} and source-dissimilar domains D_{dis}. Then in DTKD, a source-similar teacher T_{sim} and a source-dissimilar teacher T_{dis} are separately trained. Finally, we employ T_{sim} and T_{dis} to distill knowledge to a generic student S_g.

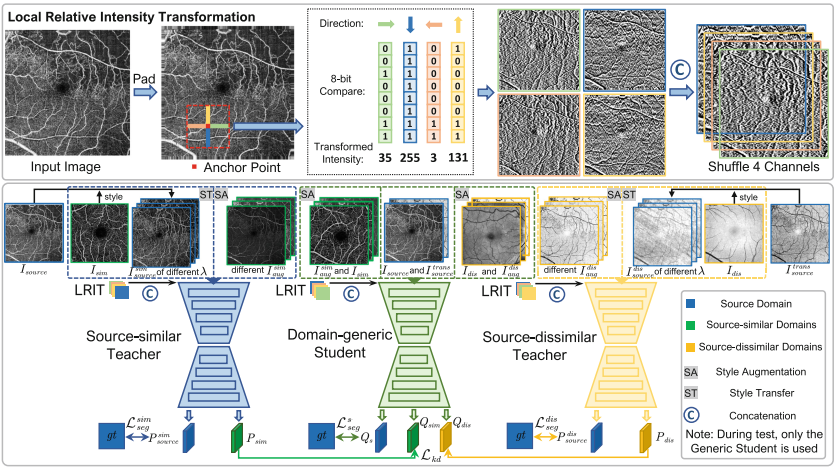


Fig. 1. Schematic demonstration of the architecture of our RVms framework. The upper part represents the LRIT module and the lower part represents the DTKD module. $P(\cdot)$ is the prediction of teachers and $Q(\cdot)$ is the prediction of the student.

2.2 Style Augmentation and Transfer

Style Augmentation. Retinal images are either grayscale or can be converted to grayscale and the content across different images is generally similar; the difference mainly lies in the intensity and the clarity of tiny vessels. Besides, the style in different modalities is also a key factor that leads to domain shift, especially for small-variation domain shift. Thus, we follow the work of [31] and utilize a non-linear transformation via the monotonic and smooth Bézier Curve function. It is a one-to-one mapping that assigns each pixel a new and unique value. We generate the Bézier Curve from two end points (P_0 and P_3) and two control points (P_1 and P_2), which is defined as

$$B(t) = (1-t)^3 P_0 + 3(1-t)^2 t P_1 + 3(1-t)t^2 P_2 + t^3 P_3, \ t \in [0,1], \qquad (1)$$

where t is a fractional value along the length of the line. We set $P_0 = (0,0)$ and $P_3 = (1,1)$ to get source-similar augmentations and the opposite to get source-dissimilar augmentations. The x-axis coordinates and y-axis coordinates of P_1 and P_2 are randomly selected from the interval $(0,1)$.

Style Transfer. In addition to grayscale differences, different modalities also differ in the image style, such as vessel shape and existence of other anatomical landmarks. To imitate the diverse appearance of different target domains, we adopt Fourier Transform to progressively translate the style of source domain to those of target domains. Given a source image x_i^s and a randomly selected target image x_j^t from target domain T_t, we first perform Fourier Transform on both images to get amplitude spectrums $\mathcal{A}_s, \mathcal{A}_t$ and phase spectrums $\mathcal{P}_s, \mathcal{P}_t$ [16,29]. Then we use a binary mask $\mathcal{M} = \mathbb{1}_{(h,w)\in[-\alpha H:\alpha H,-\alpha W:\alpha W]}$ to extract the central regions of $\mathcal{A}_s$

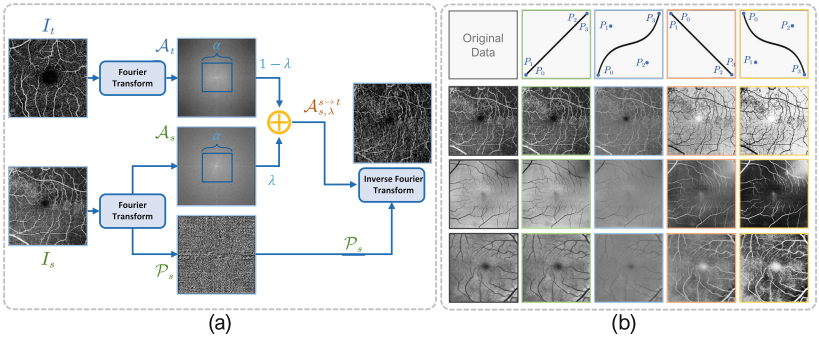


Fig. 2. Schematic demonstration of our proposed SAT module. Panel (a) is the style transfer module and Panel (b) is the style augmentation module.

and $\mathcal{A}_t$ and combine them with an interpolation rate λ. In this way the contributions of $\mathcal{A}_s$ and $\mathcal{A}_t$ to the synthesized image can be adjusted.

$$\mathcal{A}_{s,\lambda}^{s\to t} = ((1-\lambda)\mathcal{A}_s + \lambda\mathcal{A}_t) * \mathcal{M} + \mathcal{A}_s * (1-\mathcal{M}). \tag{2}$$

Finally, we perform inverse Fourier Transform to obtain an image with content from source domain and style from both source and target domains (Eq. 3). Detailed illustration of SAT is shown in Fig. 2.

$$x_{i,\lambda}^{s\to t} = \mathcal{F}^{-1}\left(\mathcal{A}_{s,\lambda}^{s\to t}, \mathcal{P}_s\right). \tag{3}$$

2.3 Dual-Teacher Knowledge Distillation

Local Relative Intensity Transformation. In our situation, despite that retinal vessels are in different modalities and multi-scaled, they always have a consistent relationship with the background in grayscale intensity. This relationship can be used to depict vessels in a domain-invariant manner. We employ LRIT to extract RV features by taking advantage of this vessel-background relationship. Inspired by [18,24], each pixel serves as an anchor point and the intensity values of the adjacent eight pixels are compared to generate a new value for the anchor point through the formula in Eq. 4. Adjacent pixels in all four directions (i.e., up, down, left, and right) are compared separately, resulting in four transformed images.

$$\begin{aligned} V_{new}(a) &= \sum_{i=1}^{8} c\left(V(a) - V\left(n_i\right)\right) \times 2^i, \\ c(x) &= \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}, \end{aligned} \tag{4}$$

where $V(\cdot)$ is the intensity value and n_i is the i-th adjacent pixel of the anchor point. For pixels near edges, we use edge padding to ensure there are eight neighboring pixels in each direction. The four transformed images are shuffled and concatenated as external channels for both teachers and student models.

Knowledge Distillation. The dual teachers T_{sim} and T_{dis} are respectively trained with augmented and style-transferred images in D_{sim} and D_{dis}. The source images, style augmented images, and style transferred images are used as inputs and the corresponding labels are used for supervision via Dice losses $\mathcal{L}_{seg}^{sim}$ and $\mathcal{L}_{seg}^{dis}$. When the two teachers are trained to converge (after τ epochs), we start the knowledge distillation process. The domain-generic student model S_g is supervised with the outputs of T_{sim} and T_{dis}, each being responsible for one type of the target domains (source-similar or source-dissimilar). Specifically, we adopt a cross-entropy loss $\mathcal{L}_{kd}$ to minimize the distribution differences between the corresponding outputs of the teachers and student models (i.e., P_{sim} and Q_{sim}, P_{dis} and Q_{dis}). The groundtruth is also used to supervise the training on the source domain through a Dice loss $\mathcal{L}_{seg}^{s}$. The total loss function is

$$\mathcal{L}_{DTKD} = \begin{cases} \mathcal{L}_{seg}^{sim} + \mathcal{L}_{seg}^{dis}, & \text{epoch } \leqslant \tau \\ \mathcal{L}_{kd} + \mathcal{L}_{seg}^{s} + \mathcal{L}_{seg}^{sim} + \mathcal{L}_{seg}^{dis}, & \text{epoch } > \tau. \end{cases} \tag{5}$$

3 Experiments and Results

3.1 Dataset

We construct a new dataset named mmRV, consisting of five domains from five publicly available datasets [1,4,13,17,25]. The details are shown in Table A1 of the appendix. For OCTA-500, we resize each image to 384 × 384 and discard samples with severe quality issues. For fundus images, we crop out the micro-vascular region surrounding the macula to avoid interference from other unrelated structures such as the optic disc and only focus on vessels; the capillaries near the macula are mostly multi-scaled and difficult to segment [15]. We also apply Contrast Limited Adaptive Histogram Equalization to fundus images as preprocessing. For PRIME-FP20, we augment each image for four times.

3.2 Experimental Setting

We train our RVms framework on the newly-constructed mmRV dataset. We conduct two sets of MTDA experiments respectively using DRIVE and OCTA-500 as the source domain because the image modalities in those two datasets are the most commonly used in clinical practice and those two datasets are most well-annotated. All compared methods and RVms are implemented with Pytorch using NVIDIA RTX 3090 GPUs. For both teachers and student models in RVms, we use the Adam optimizer with a learning rate of 1×10^{-3}. The number of epochs τ for training the dual teachers is set to be 200 and then we co-train the teachers with the student for another 400 epochs. α is set to 0.2 and λ is randomly selected in $(0, 1)$. During testing, the target domain images are directly inputted to S_g to get the corresponding predictions. We use ResNet34 [7] with ImageNet pretrained initialization as the encoder in a modified U-net [21] architecture.

Table 1. Quantitative evaluations of different methods. **Bold** and underlined numbers respectively denote the best and second-best results.

	Source domain: OCTA (OCTA-500)											
Modalities	Fundus image (DRIVE)		OCTA (ROSE)		OCT (OCTA-500)		Fundus image (HRF)		UWF fundus (PRIME-FP20)		Average	
Metrics	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓
Source only	2.18	178.06	52.35	25.52	5.70	62.89	0.43	189.60	1.33	180.42	12.40	127.30
Dofe	24.03	27.91	48.52	21.21	34.41	21.76	24.67	25.63	30.37	24.75	32.40	24.25
ADVENT	1.46	52.53	52.86	12.80	4.91	44.48	0.37	44.16	1.01	20.11	12.12	34.81
Multi-Dis	1.45	39.65	58.98	**10.29**	5.07	30.77	0.31	46.91	1.03	52.09	13.37	35.94
T_{sim}	5.35	36.18	53.04	23.27	2.92	32.89	2.90	30.75	5.07	37.08	13.85	39.27
T_{dis}	**72.77**	**10.77**	5.68	19.52	**77.76**	**15.60**	72.90	**12.12**	**76.49**	**12.57**	49.73	**14.16**
Ours	72.64	17.72	**60.80**	13.49	72.81	24.70	**74.03**	12.97	74.54	21.09	**70.96**	17.99
Oracle	67.72	9.05	71.43	12.14	82.68	11.13	72.09	12.67	77.80	9.33	74.34	10.86
	Source domain: fundus image (DRIVE)											
Modalities	OCTA (OCTA-500)		OCTA (ROSE)		OCT (OCTA-500)		Fundus image (HRF)		UWF fundus (PRIME-FP20)		Average	
Metrics	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓
Source only	11.39	21.81	17.69	**10.24**	59.24	19.77	58.77	**10.38**	64.87	14.92	42.39	**15.42**
Dofe	30.94	22.72	38.39	11.48	62.69	19.13	60.28	12.86	65.46	**13.88**	51.55	16.01
ADVENT	22.38	47.40	20.36	33.88	47.32	42.66	53.15	34.59	52.86	42.68	39.21	40.24
Multi-Dis	32.75	21.30	33.73	11.16	52.42	35.79	59.46	28.22	60.62	38.54	47.80	27.00
T_{sim}	12.05	19.58	14.32	23.08	59.11	21.19	59.15	15.37	64.96	17.69	41.91	19.38
T_{dis}	64.44	18.51	**63.81**	16.13	11.01	**14.63**	6.87	25.28	10.08	24.69	31.24	19.84
Ours	**68.93**	**15.67**	61.31	16.05	**64.35**	18.93	**63.18**	13.50	**69.44**	14.04	**65.44**	15.63
Oracle	88.35	4.21	71.43	12.14	82.68	11.13	72.09	12.67	77.80	9.33	78.47	9.89

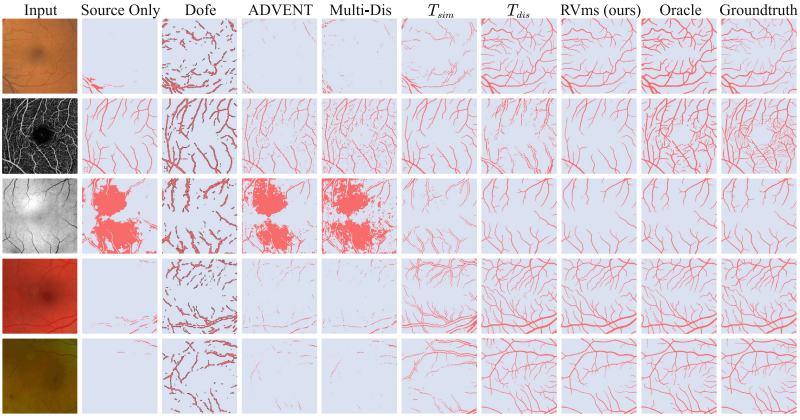


Fig. 3. Representative RV segmentation results using the OCTA images from OCTA-500 as the source domain. The target domains are Fundus Image (DRIVE), OCTA (ROSE), OCT (OCTA-500), Fundus Image (HRF), and UWF Fundus (PRIME-FP20) from top to bottom.

3.3 Results

All methods are evaluated using two metrics, i.e., Dice[%] and 95% Hausdorff Distance (HD[px]), the results of which are tabulated in Table 1. We compare RVms with two recently-developed SOTA DA/MTDA models, namely ADVENT [26] and Multi-Dis [23]. Note that [26] is trained with mixed target domains. We also compare with a domain generalization method Dofe [27], which leaves one out as the target domain and uses multiple source-similar domains as the source domains. The results from the dual teachers T_{sim} and T_{dis} are also reported. **Source Only** means the model is trained with source domain data only. **Oracle** means the model is trained and tested on the specific target domain. Our method achieves Dice scores that are about 38% and 13% higher than the second best method when using OCTA (OCTA-500) and fundus image (DRIVE) as the source domain. It is evident that our proposed RVms delivers superior RV segmentation performance when encountering both cross-modality and cross-center domain shift. Besides, it achieves the highest average Dice score.

Representative visualization results are illustrated in Fig. 3 and Fig. A1 of the appendix. We observe that all compared methods fail in many cases while our framework is very close to the Oracle. Apparently, our framework achieves superior performance when tested on all target domains.

To evaluate the effectiveness of several key components in RVms, we conduct ablation studies. We compare with the proposed RVms without style augmentation (w/o sa), without style transfer (w/o st), without LRIT (w/o LRIT). We also compare with training a single model with images in D_{sim} and D_{dis} without knowledge distillation (w/o KD) and employing only one teacher model (w/o DT). The results are shown in Table 2 and Table A2 of the appendix. The performance degrades when removing any component in the framework, in terms of the average Dice score.

Table 2. Ablation analysis results for several key components in our proposed framework using OCTA (OCTA-500) as the source domain. **Bold** and underlined numbers respectively denote the best and second-best results.

	Source domain: OCTA (OCTA-500)											
Modalities	Fundus image (DRIVE)		OCTA (ROSE)		OCT (OCTA-500)		Fundus image (HRF)		UWF fundus (PRIME-FP20)		Average	
Metrics	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓
Source only	2.18	178.06	52.35	25.52	5.70	62.89	0.43	189.60	1.33	180.42	12.40	127.30
w/o sa	62.74	24.82	48.25	26.51	60.47	30.50	60.48	30.89	60.17	29.91	58.42	26.73
w/o st	70.19	21.58	41.95	27.29	71.59	24.81	70.86	18.19	69.57	23.33	55.94	23.04
w/o LRIT	71.86	**13.35**	56.65	19.93	71.60	18.84	71.54	**10.28**	73.83	**15.17**	69.09	**15.31**
w/o KD	69.73	14.78	13.08	20.57	71.07	**18.83**	70.53	16.75	71.96	18.54	59.27	17.89
w/o DT	70.08	16.06	43.20	26.12	72.11	20.91	72.21	13.85	72.25	19.64	65.97	19.31
Ours	**72.64**	17.72	**60.80**	**13.49**	**72.81**	24.70	**74.03**	12.97	**74.54**	21.09	**70.96**	17.99
Oracle	67.72	9.05	71.43	12.14	82.68	11.13	72.09	12.67	77.80	9.33	74.34	10.86

4 Conclusion

In this paper, we proposed and validated a novel framework for unsupervised multi-target domain adaptation in retinal vessel segmentation. We used style augmentation and style transfer to generate source-similar images and source-dissimilar images to improve the robustness of both teachers and student models. We also conducted knowledge distillation from dual teachers to a generic student, wherein a domain invariant method named LRIT was utilized to facilitate the training process. Another contribution of this work is that we constructed a new dataset called mmRV from several public datasets, which can be used as a new benchmark for DA and domain generalization. Through extensive experiments, our proposed RVms was found to largely outperform representative SOTA MTDA methods, in terms of RV segmentation from different modalities.

Acknowledgements. This study was supported by the Shenzhen Basic Research Program (JCYJ20190809120205578); the National Natural Science Foundation of China (62071210); the Shenzhen Science and Technology Program (RCYX20210609103056042); the Shenzhen Basic Research Program (JCYJ20200925153847004); the Shenzhen Science and Technology Innovation Committee (KCXFZ2020122117340001).

References

1. Budai, A., Bock, R., Maier, A., et al.: Robust vessel segmentation in fundus images. Int. J. Biomed. Imaging **2013**(6), 154860 (2013)
2. Cai, J., Zhang, Z., Cui, L., et al.: Towards cross-modal organ translation and segmentation: a cycle-and shape-consistent generative adversarial network. Med. Image Anal. **52**, 174–184 (2019)
3. Cheng, P., Lin, L., Huang, Y., Lyu, J., Tang, X.: I-secret: importance-guided fundus image enhancement via semi-supervised contrastive constraining. In: Medical Image Computing and Computer Assisted Intervention, vol. 12908, pp. 87–96. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_9
4. Ding, L., Kuriyan, A.E., Ramchandran, R.S., et al.: Weakly-supervised vessel detection in ultra-widefield fundus photography via iterative multi-modal registration and learning. IEEE Trans. Med. Imaging **40**(10), 2748–2758 (2020)
5. Dou, Q., et al.: PnP-AdaNet: plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. IEEE Access **7**, 99065–99076 (2019)
6. Ganin, Y., Ustinova, E., Ajakan, H., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2030–2096 (2016)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
9. Hu, Z., Liu, Q., Paulus, Y.M.: New frontiers in retinal imaging. Int. J. Ophthalmic Res. **2**(3), 148–158 (2016)
10. Huang, D., Swanson, E.A., Lin, C.P., et al.: Optical coherence tomography. Science **254**(5035), 1178 (1991)

11. Huang, Y., Lin, L., Li, M., et al.: Automated hemorrhage detection from coarsely annotated fundus images in diabetic retinopathy. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1369–1372. IEEE (2020)
12. Javanmardi, M., Tasdizen, T.: Domain adaptation for biomedical image segmentation using adversarial training. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 554–558. IEEE (2018)
13. Li, M., Zhang, Y., Ji, Z., et al.: IPN-V2 and OCTA-500: methodology and dataset for retinal image segmentation. arXiv preprint arXiv:2012.07261 (2020)
14. Lin, L., Wang, Z., Wu, J., et al.: BSDA-Net: a boundary shape and distance aware joint learning framework for segmenting and classifying octa images. In: Medical Image Computing and Computer Assisted Intervention, vol. 12908, pp. 65–75. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_7
15. Lin, L., Wu, J., Cheng, P., Wang, K., Tang, X.: BLU-GAN: bi-directional ConvLSTM U-net with generative adversarial training for retinal vessel segmentation. In: Intelligent Computing and Block Chain, vol. 1385, pp. 3–13. Springer, Cham (2020). https://doi.org/10.1007/978-981-16-1160-5_1
16. Liu, Q., Chen, C., Qin, J., et al.: FedDG: federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1013–1023 (2021)
17. Ma, Y., Hao, H., Xie, J., et al.: Rose: a retinal oct-angiography vessel segmentation dataset and new model. IEEE Trans. Med. Imaging **40**(3), 928–939 (2020)
18. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
19. Peng, L., Lin, L., Cheng, P., Huang, Z., Tang, X.: Unsupervised domain adaptation for cross-modality retinal vessel segmentation via disentangling representation style transfer and collaborative consistency learning. arXiv preprint arXiv:2201.04812 (2022)
20. Peng, L., Lin, L., Cheng, P., et al.: Fargo: a joint framework for FAZ and RV segmentation from octa images. In: Fu, H., Garvin, M.K., MacGillivray, T., Xu, Y., Zheng, Y. (eds.) Ophthalmic Medical Image Analysis, vol. 12970, pp. 42–51. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87000-3_5
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds.) Medical Image Computing and Computer-Assisted Intervention, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
22. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision, vol. 6314, pp. 213–226. Springer, Cham (2010). https://doi.org/10.1007/978-3-642-15561-1_16
23. Saporta, A., Vu, T.H., Cord, M., Pérez, P.: Multi-target adversarial frameworks for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9072–9081 (2021)
24. Shi, T., Boutry, N., Xu, Y., Géraud, T.: Local intensity order transformation for robust curvilinear object segmentation. IEEE Trans. Image Processing **31**, 2557–2569 (2022)
25. Staal, J., Abràmoff, M.D., Niemeijer, M., et al.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**(4), 501–509 (2004)

26. Vu, T.H., Jain, H., et al.: Advent: adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2517–2526 (2019)
27. Wang, S., Yu, L., Li, K., et al.: DoFE: domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. IEEE Trans. Med. Imaging **39**(12), 4237–4248 (2020)
28. Wang, S., Yu, L., Yang, X., et al.: Patch-based output space adversarial learning for joint optic disc and cup segmentation. IEEE Trans. Med. Imaging **38**(11), 2485–2495 (2019)
29. Yang, Y., Soatto, S.: FDA: fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4085–4095 (2020)
30. Zhang, T., Cheng, J., et al.: Noise adaptation generative adversarial network for medical image analysis. IEEE Trans. Med. Imaging **39**(4), 1149–1159 (2019)
31. Zhou, Z., Sodha, V., Rahman Siddiquee, M.M., et al.: Models genesis: generic autodidactic models for 3D medical image analysis. In: Medical Image Computing and Computer Assisted Intervention, vol. 11767, pp. 384–393. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_42
32. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2223–2232 (2017)



Rethinking Degradation: Radiograph Super-Resolution via AID-SRGAN

Yongsong Huang[1(✉)], Qingzhong Wang[2], and Shinichiro Omachi[1]

[1] Tohoku University, Sendai, Japan
huang.yongsong.r1@dc.tohoku.ac.jp, machi@ecei.tohoku.ac.jp
[2] Baidu Research, Beijing, China
wangqingzhong@baidu.com

Abstract. In this paper, we present a medical **A**ttent**I**on **D**enoising **S**uper **R**esolution **G**enerative **A**dversarial **N**etwork (AID-SRGAN) for diographic image super-resolution. First, we present a medical practical degradation model that considers various degradation factors beyond downsampling. To the best of our knowledge, this is the first composite degradation model proposed for radiographic images. Furthermore, we propose AID-SRGAN, which can simultaneously denoise and generate high-resolution (HR) radiographs. In this model, we introduce an attention mechanism into the denoising module to make it more robust to complicated degradation. Finally, the SR module reconstructs the HR radiographs using the "clean" low-resolution (LR) radiographs. In addition, we propose a separate-joint training approach to train the model, and extensive experiments are conducted to show that the proposed method is superior to its counterparts. *e.g.*, our proposed method achieves 31.90 of PSNR with a scale factor of 4×, which is 7.05% higher than that obtained by recent work, SPSR [16]. Our dataset and code will be made available at: https://github.com/yongsongH/AIDSRGAN-MICCAI2022.

Keywords: Musculoskeletal radiographs · Super-resolution

1 Introduction

High-resolution musculoskeletal radiographs provide more details that are crucial for medical diagnosis, particularly for diagnosing primary bone tumors and bone stress injuries [2,4,8,18,22]. However, radiographic image quality is affected by many factors, such as scanning time, patients' poses, and motions, and achieving higher-resolution medical images is expensive and time-consuming because it requires a relatively long scanning time. However, existing SR algorithms fail to fully consider the degradation factors mentioned above.

Imperfect degenerate models put the algorithm at risk of domain shift (see Table 2: (a) Domain shift). To solve this problem, we first need to rethink the degradation – high quality transforms into lower quality. The degradation of radiographs is related to statistical noise, external disruptions, and downsampling [17,27,30,31]. First, the two basic types of statistical noise – Poisson and

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 43–52, 2022.
https://doi.org/10.1007/978-3-031-21014-3_5

Gaussian-are common in radiographic images [9,27]. However, most of the existing state-of-the-art deep learning-based SR methods focus only on the degradation of bicubic downsampling [12,16,21], *i.e.*, the models directly take the bicubic downsampling images as input and reconstruct the HR images, leading to the problem of domain shift when applied to noisy images. Second, the disruptions in the application scenarios arise from the following factors: radiologists and patients, as well as information loss due to compressed transmission via the Internet [6,19]. In general, operational mistakes and the patients' displacement relative to the device would introduce motion blur [1,21]. On the other hand, telemedicine [10] requires uploading medical images to a data center for easy consultation and online storage. Owing to the limited network bandwidth, compression [6] is widely employed for online data transmission, leading to low-quality radiographic images.

To address this problem, more attention has been paid to the tasks of medical image super-resolution [7,18,25]. Deep learning-based methods [12,16,29,33,34] dominate image SR, which learns a mapping from LR images to HR images and differs from traditional methods in which more prior knowledge is required [3]. Recently, blind SR that considers real-world degradation has drawn much attention [28,30,33,34] because it is more common. However, only Gaussian blur and compression were adopted to synthesize the paired training data [30] and the parameters of the Gaussian kernels were fixed [33,34]. In fact, the degradation of radiographs could be more complicated, including statistical noise, motion blur, and compression. In this study, we aim to address the problem of musculoskeletal radiograph SR. The main contributions are in three-fold.

- We propose a practical degradation model for radiographs, which considers most possible degradation factors, such as statistical noise, motion blur, compression, and each of them has variant parameters. This model aims to represent complex nonlinear degeneracies, unlike current models that focus more on downsampling. In addition, the degradation model is applied to synthesize data to train the proposed SR model.
- We propose a medical attention denoising SRGAN model (AID-SRGAN). An attention mechanism is introduced into the denoising module to make it more robust to complicated degradation. Moreover, we propose a two-stage training approach to train the proposed AID-SRGAN, *i.e.*, we first separately train the denoising module and SR module to obtain a relatively good denoising network and SR network, respectively. We then jointly train the denoising and SR modules in an end-to-end manner to further improve the performance. Finally, it is a flexible framework and easy to follow.
- We conduct extensive experiments and compare the proposed model with other existing works. AID-SRGAN is superior to its counterparts, achieving 31.90 dB of PSNR with a scale factor of 4×, which is 7.05% higher than a recent work – SPSR [16]. Moreover, considering SSIM, AID-SRGAN outperforms SPSR, *e.g.*, 0.9476 vs. 0.9415. In addition, ablation studies on the modules and hyper-parameters are conducted to demonstrate their effectiveness.

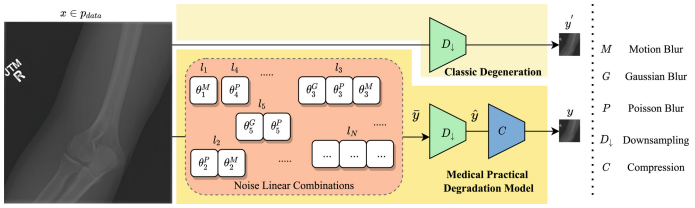


Fig. 1. Overview of the medical practical degradation model. The model takes x – HR radiographs as input and imposes L_N – noise linear combinations on x, then bicubic downsampling and compression are applied to the noisy radiographs, yielding noisy LR radiographs y. We also generate LR radiographs y' via directly imposing downsampling on HR radiographs.

2 Methodology

2.1 Medical Practical Degradation Model

In the real world, HR images x are degraded to LR images y, with random and complicated degradation mechanisms [28,30]. An HR image suffers from blur kernels k, such as statistical noise, which first degrades the image quality. Furthermore, the damaged image $\bar{y}$ is transformed into an LR image $\hat{y}$ by downsampling $D_\downarrow$. Finally, image compression C is used for online transmission and storage. The entire procedure is represented using the following function:

$$y = \mathcal{F}^D(x) = C\left(D_\downarrow(x \oplus k)\right), \tag{1}$$

where $\mathcal{F}^D$ denotes a degradation function. However, existing SR models only consider one or two degradation factors [12,16,29,34], which could result in a large gap between open world LR images and synthetic LR images [30] and lead to poor performance in practice due to the domain-shift problem.

In this study, we propose a practical degradation model for radiographs (Fig. 1), where we consider most degradation factors in radiographs – statistical noise combinations L_N, downsampling $D_\downarrow$, and compression C. In L_N, we combine the Gaussian blur G, Poisson blur P and motion blur M, which are widely observed in radiographs. In particular, the statistical noise parameters θ^{L_N}, such as expectation μ and variance $\mathbf{\Sigma}$, are automatically updated when a new sample x is fed into the degradation model, yielding a noisy image $\bar{y}$, *i.e.*,

$$\bar{y} = L_N\left(x; \theta^{L_n}\right) \quad \theta^{L_N} \sim p\left(\theta^{L_N}\right). \tag{2}$$

Then, $\bar{y}$ is resized using downsampling and compressed to generate noisy LR images y.

From the perspective of latent space [14], the proposed medical practical degradation model applies external variables θ and latent variables $\bar{y}, \hat{y}$ to approximate the real-world degradation distribution $p(y|x)$, *i.e.*, $\tilde{p}(y \mid x) \approx p(y \mid$

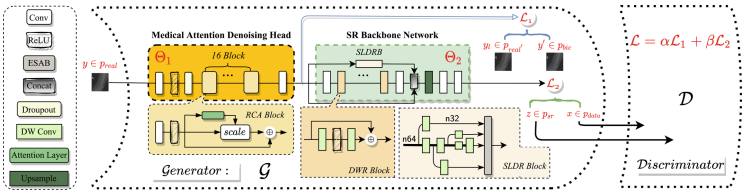


Fig. 2. Overview of our proposed AID-SRGAN. We can jointly train medical attention denoising head and SR backbone network in an end-to-end manner. AID-SRGAN with the input $y \in p_{real}$, HR image $x \in p_{data}$ and $z \in p_{sr}$ – the output fake sample by generator $\mathcal{G}$. $\mathcal{D}$ is the discriminator. $y' \in p_{bic}$ denotes the sample obtained using downsampling degradation only. $\mathcal{L}$ is a loss function. Θ_1 and Θ_2 represent the parameters of $\mathcal{G}$. α and β are hyperparameters. Zoom in for best view.

x), where y and x are LR and HR images, respectively. The distribution can be computed as follows:

$$\tilde{p}(y \mid x) = \int p(y \mid \hat{y}, \theta) p(\hat{y} \mid \bar{y}, \theta) p(\bar{y} \mid x, \theta) p(\theta) d\hat{y} d\bar{y} d\theta. \tag{3}$$

We draw samples y from $\tilde{p}(y|x)$ and build LR-HR pairs $\{(y_1, x_1), \cdots, (y_n, x_n)\}$ to train AID-SRGAN. This degradation model is expected to attract researchers to focus on medical image sample representation more.

2.2 Medical Attention Denoising SRGAN (AID-SRGAN)

For AID-SRGAN, our goal is to propose a straightforward model that is easy to follow. After obtaining the training data, there are different approaches to solve the real-world super-resolution problem, such as 1) denoising first and then reconstructing HR images, or 2) direct reconstruction [28,30,33,34]. The proposed AID-SRGAN adopts denoising first and then reconstructs the HR images. Moreover, denoising first sounds more reasonable because LR images are noisy. We demonstrate the framework of the AID-SRGAN in Fig. 2, which is composed of two modules: denoising, which reconstructs the LR images obtained by downsampling only, and super-resolution, which reconstructs the HR image from the reconstructed "clean" LR image. In summary, the denoising module takes noisy LR images y as input and reconstructs y', while the SR module takes y' as input and reconstructs HR images x. For training strategies, we can separately train the two modules to obtain a relatively good initialization of AID-SRGAN and then train the two modules in an end-to-end manner via backpropagation.

Medical Attention Denoising: one observation of the denoising deep neural networks is that the activation map varies using different degradation factors. Figure 3 presents the activation map with different degradation kernels, and it can be observed that the network pays more attention to objects, such as characters using motion blur, while focusing on the entire image using Gaussian blur.

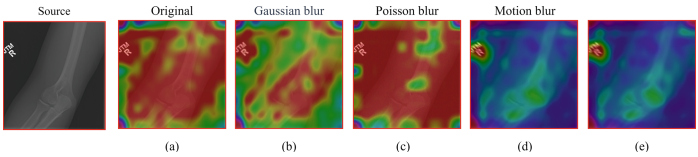


Fig. 3. Visualization of the heat map for the degradation representation with different factors by CAM [23]. (e) illustrates representations generated by motion blur, Gaussian blur, and compression. It is necessary to guide the denoising, considering the model's different response to various degradation factors. Best viewed in color. (Color figure online)

In summary, there will be different responses for different degradation factors. To obtain consistent representations and adapt to different degradation factors, we introduce an attention mechanism [11] to guide the denoising procedure. We adopted the residual channel attention (RCA)block as the basic unit for the attention denoising head, which is calculated as follows:

$$x^{l+1} = x^l + \sigma\left(Conv(x^l)\right) \odot x^l, \tag{4}$$

where x^l denotes the lth layer input, $Conv(\cdot)$ represents the convolution layer, $\sigma(\cdot)$ denotes the activation function, and $\odot$ represents the element-wise multiplication. In contrast to other methods, our network does not estimate the blur kernel, which is beneficial for reducing reliance on extensive prior knowledge [26].

SR Backbone Network: When the distribution of y_l is available, we seek to reconstruct $z \in p_{sr}$ from $y_l \in p_{real'}$ using an adversarial training approach. We also observed that PSRGAN [12] performs reliably in image reconstruction for images with simple patterns and structures, such as infrared and gray images, which benefits radiological image reconstruction.

$$\begin{aligned} &\min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{D}, \mathcal{G}) = E_{x \sim p_{\text{data}}} \left[\log \mathcal{D}(x)\right] \\ &+ E_{y_l \sim p_{\text{real}'}, y \sim p_{\text{real}}} \left[\log\left(1 - \mathcal{D}\left(z = \mathcal{G}\left(y_l \mid y\right)\right)\right)\right]. \end{aligned} \tag{5}$$

For the generator $\mathcal{G}$, the output (from the denoising head) is fed to the SR backbone network. The backbone network consists of the main and branch paths, which are built using DWRB and SLDRB, respectively. The detailed information is shown in Fig. 2, where the key component-SLDRB in the branch is introduced the information distillation. This distillation method improves the feature representation of the model by setting different numbers of feature channels (as shown in Fig. 2: n64 & n32), which is believed to be beneficial for images with fewer patterns in the experiments [12]. The discriminator $\mathcal{D}$ is trained to maximize the probability of providing a real sample to both training data and fake samples generated from $\mathcal{G}$. $\mathcal{G}$ is trained to minimize $\log\left(1 - \mathcal{D}\left(z = \mathcal{G}\left(y_l \mid y\right)\right)\right)$, where $y_l \in p_{real'}$ denotes the bicubic LR input. The objective function is defined by Eq. 5, where V denotes the divergence. E represents cross-entropy.

Table 1. PSNR↑ and SSIM↑ results of different methods on MURA-mini (mini)& MURA-plus (plus) with scale factors of 4 & 2. Ours and Ours+ have 16 and 256 RCA blocks, respectively. There are two test datasets with the same HR images but different degraded LR images (which are more damaged). The best results are in **bold**.

			Ours+	Ours	Bic	DPSR	PSRGAN	SRMD	ESRGAN	SPSR
×4	mini	PSNR	**31.90**	31.21	28.55	29.79	29.39	16.32	28.63	29.80
		SSIM	0.9476	**0.9506**	0.9403	0.9497	0.9306	0.6247	0.9313	0.9415
	plus	PSNR	**31.52**	30.78	28.68	29.98	28.65	16.31	28.07	28.77
		SSIM	0.9454	**0.9469**	0.9354	0.9431	0.9273	0.6204	0.9267	0.9358
×2	mini	PSNR	34.03	**34.11**	29.91	31.00	29.87	14.75	30.03	30.74
		SSIM	0.9589	**0.9590**	0.9440	0.9412	0.9470	0.5571	0.9438	0.9357
	plus	PSNR	**32.54**	32.43	30.21	31.59	29.04	14.74	29.43	30.53
		SSIM	0.9473	**0.9483**	0.9361	0.9322	0.9381	0.5570	0.9340	0.9313

2.3 Training Strategy

We first separately train the denoising network and SR network using the synthetic data, *i.e.*, the paired data (y, y') is employed to train the denoising network, and (y', x) is used to train the SR network. In summary, we seek to determine the optimal parameters θ^* by minimizing the expected risk.

$$\theta^* = \underset{\theta \in \Theta_1}{\operatorname{argmin}} \mathbf{E}_{y,y'} \left[\mathcal{L}_1 \left(y, y', \theta \right) \right] \tag{6}$$

where $\mathcal{L}_1 \left(y, y', \theta \right)$ is a loss function that depends on parameter θ. y and y' denote the input image and bicubic downsampling image, respectively. After separate training, we employ (y, x) to jointly train the entire framework in an end-to-end manner. Separate training obtains a good local optimal of AID-SRGAN, and joint training further boosts the performance, which is similar to the pre-training-fine-tuning paradigm [5]; however, we employ supervised pre-training. According to the pre-experimental results, the medical attention denoising head performs better when compared with the direct use of DNCNN denoising or DNN-based denoising heads (see Table 2: (b) Network selection & (c) RCA Block).

3 Experiments

Dataset: we employed a widely used dataset, MURA [20] to synthesize training pairs. MURA contains 40,005 musculoskeletal radiographs of the upper extremities. We selected 4,000 images as the training set, named MURA-SR. Two test datasets were used, MURA-mini and MURA-plus. Both were composed of 100 HR images and different degraded LR images. For MURA-SR, the blur kernel size was randomly selected from $\{1, 3, ..., 11\}$, whereas MURA-mini and MURA-plus used kernel sizes selected from $\{1, 3, 5\}$ and $\{7, 9, 11\}$,respectively, *i.e.*, the

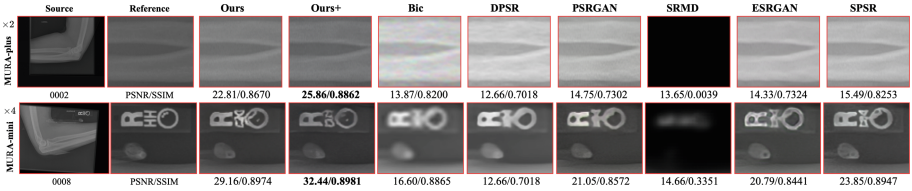


Fig. 4. Top: Qualitative comparisons on 002 samples from MURA-plus with upsampling scale factor of 2. Bottom: Qualitative comparisons on 008 samples from MURA-mini with upsampling scale factor of 4. The PSNR↑/SSIM↑ of the test images are shown in the figure. The best results are in **bold**. Zoom in for best view.

degradation of MURA-plus is more serious. The probability of using a blur kernel was randomly selected from $\{0.1, 0.2, ..., 1.0\}$. Finally, the JPEG compression quality factor was set to 3 [15,30] for all images.

Training Details: We trained our model with a batch size of 32 on two TITAN X (Pascal) GPUs. The training HR patch size was set to 96. We employed the Adam optimizer [13] with a learning rate of $1e-5$. We used VGG16 [24] as the discriminator and the combination of L1 and SSIM as the loss function [30].

3.1 Results

Quantitative Results: in experiments, we propose two versions of AID-SRGAN: *Ours* (16 RCA Blocks) and *Ours+* (256 RCA Blocks). As shown in Table 1, we compare our approach with downsampling-based models[1] (PSRGAN [12], ESRGAN [29],SPSR [16]), and real-world oriented models[2] (DPSR [34], SRMD [33]), and we trained all the models using MURA-SR. The metric scores (PSNR and SSIM) were calculated for the Y channel of the YCbCr space. We can easily conclude that our proposed model is superior to the counterparts on both MURA-mini and MURA-plus with the upsampling factors of 4 and 2 considering PSNR and SSIM, *e.g.*, Ours+ achieves 31.90 dB of PSNR on MURA-mini with the scale factor of 4, which is roughly 7.08% higher than that achieved by DPSR. Compared with downsampling-based methods (i.e., PSRGAN and ESRGAN)our proposed approach achieves a more remarkable performance, and the average relative improvement is approximately 10%, indicating that denoising matters in radiograph super-resolution. When comparing Ours+ and Ours, it is evident that deeper denoising modules achieve a higher PSNR, whereas shallow denoising networks obtain a slightly higher SSIM. A possible reason is that shallow networks converge faster than deep networks, and we use SSIM in the loss function; hence, shallow networks can reach a higher SSIM score with the same training epochs.

Qualitative Results: Figure 4 presents some examples of the reconstructed HR radiographs generated by different models. Compared with the existing models,

[1] The degradation model only consider downsapling.
[2] The degradation model considers downsampling and others, such as Gaussian blur.

Table 2. Ablation on using more solutions (network structure and test dataset). SR network is PSRGAN. The denoising model is DNCNN. More setting details as follows: upsampling factor: ×4, evaluated in RGB space.

(a) Domain shift.

Test dataset	**PSNR**/*dB*	**SSIM**↑
Bicubic	33.66	0.9383
MURA-mini	26.85	0.8445

(b) Network selection.

Modules ablation	**PSNR**/*dB*	**SSIM**↑
Direct. SR	28.05	0.9038
Denoising+SR	28.09	0.9072

(c) RCA Block.

Denoising head	**PSNR**/*dB*	**SSIM**↑
DNCNN	27.34	0.8747
+Att.	28.78	0.9250

(d) Hyperparameters.

Metrics	+dropout (P=0.5)	+dropout (P=0.1)	+denoising pretrain α = 1e-5/β=0	+SR pretrain β = 1e-5/α=0	+joint train β = 1e-5/α=1e-5
PSNR/*dB*	29.02	29.34	29.13	29.90	30.14
SSIM↑	0.9324	0.9179	0.9236	0.9281	0.9271

the proposed AID-SRGAN can generate a sharper HR image and reconstruct more details, *e.g.*, the edge of the bone in the first image, and the characters in the second image, and the PSNR and SSIM scores are also relatively high compared to the counterparts.

3.2 Ablation Studies

We conducted extensive ablation studies on different modules and hyperparameters in the AID-SRGAN, and the comparison is presented in Table 2. In complex degenerate models, the SR algorithm will have difficulties fitting the data distribution. This also explains the domain drift (see (a), domain shift), and paying more attention to the degradation model is beneficial. In (b), network selection, we can see that SR+ denoising directly outperforms SR using DNCNN [32] as the denoising network. The performance can be boosted to 28.78 dB using the RCA block, as shown in (c), while the ablation experiments with hyperparameters are shown in (d). Using dropout, we can further improve the performance, achieving 29.34 dB of PSNR. Finally, we jointly trained the denoising and SR networks based on the separate pre-training models, achieving 30.14 of PSNR and 0.9271 of SSIM.

4 Conclusion

In this study, we presented the AID-SRGAN model for musculoskeletal radiograph super-resolution. In addition, we introduced residual channel attention (RCA) block for complex degradation factors. To train the proposed model and adapt to the open world SR task, we further proposed a medical degradation model that included most possible degradation factors, such as Gaussian blur, motion blur, and compression. Finally, the experimental results show that the proposed model outperforms its counterparts in terms of PSNR and SSIM. In the future, further studies will be carried out to validate degradation models for other medical images.

References

1. Asli, H.S., et al.: Motion blur invariant for estimating motion parameters of medical ultrasound images. Sci. Rep. **11**(1), 1–13 (2021)
2. Beatriz, M., et al.: Using super-resolution generative adversarial network models and transfer learning to obtain high resolution digital periapical radiographs. Comput. Biol. Med. **129**, 104139 (2021)
3. Chen, H., et al.: Real-world single image super-resolution: a brief review. Inf. Fusion **79**, 124–145 (2022)
4. Christ, A.B., et al.: Compliant compression reconstruction of the proximal femur is durable despite minimal bone formation in the compression segment. Clin. Orthop. Relat. Res.® **479**(7), 1577–1585 (2021)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Dimililer, K.: DCT-based medical image compression using machine learning. Sig. Image Video Process. **16**, 55–62 (2022). https://doi.org/10.1007/s11760-021-01951-0
7. de Farias, E.C., Di Noia, C., Han, C., Sala, E., Castelli, M., Rundo, L.: Impact of GAN-based lesion-focused medical image super-resolution on the robustness of radiomic features. Sci. Rep. **11**(1), 1–12 (2021)
8. Groot, O.Q., et al.: Does artificial intelligence outperform natural intelligence in interpreting musculoskeletal radiological studies? A systematic review. Clin. Orthop. Relat. Res. **478**(12), 2751 (2020)
9. Guan, M., et al.: Perceptual quality assessment of chest radiograph. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 Proceedings, Part VII. vol. 12907, pp. 315–324. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87234-2_30
10. Heller, T., et al.: Educational content and acceptability of training using mobile instant messaging in large HIV clinics in Malawi. Ann. Global Health **87** (2021). https://doi.org/10.5334/aogh.3208
11. Hu, J., et al.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. **42**, 2011–2023 (2020)
12. Huang, Y., et al.: Infrared image super-resolution via transfer learning and PSRGAN. IEEE Sign. Process. Lett. **28**, 982–986 (2021)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Lee, S., Ahn, S., Yoon, K.: Learning multiple probabilistic degradation generators for unsupervised real world image super resolution. ArXiv abs/2201.10747 arXiv:abs/2201.10747 (2022)
15. Liu, D., et al.: Non-local recurrent network for image restoration. Adv. Neural Inf. Process. Syst. **31** (2018)
16. Ma, C., et al.: Structure-preserving image super-resolution. IEEE Trans. Pattern Anal. Mach. Intell. 1 (2021). https://doi.org/10.1109/TPAMI.2021.3114428
17. Mohan, K.A., Panas, R.M., Cuadra, J.A.: Saber: a systems approach to blur estimation and reduction in x-ray imaging. IEEE Trans. Image Process. **29**, 7751–7764 (2020)

18. Peng, C., Zhou, S.K., Chellappa, R.: DA-VSR: domain adaptable volumetric super-resolution for medical images. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2021. Lecture Notes in Computer Science, vol. 12906, pp. 75–85. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_8
19. Peng, H., et al.: Secure and traceable image transmission scheme based on semitensor product compressed sensing in telemedicine system. IEEE Internet Things J. **7**(3), 2432–2451 (2020)
20. Rajpurkar, P., et al.: Mura: large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957 (2017)
21. Rezaei, M., Yang, H., Meinel, C.: Deep learning for medical image analysis. ArXiv abs/1708.08987 arXiv:1708.08987 (2017)
22. von Schacky, C.E., et al.: Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. Radiology **301**(2), 398–406 (2021)
23. Selvaraju, R.R., et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. **128**, 336–359 (2019)
24. Simonyan, et al.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. van Sloun, R.J., et al.: Super-resolution ultrasound localization microscopy through deep learning. IEEE Trans. Med. Imaging **40**(3), 829–839 (2020)
26. Son, S., et al.: Toward real-world super-resolution via adaptive downsampling models. IEEE Trans. Pattern Anal. Mach. Intell. (2021). https://doi.org/10.1109/TPAMI.2021.3106790
27. Thanh, D.N.H., et al.: A review on CT and x-ray images denoising methods. Informatica (Slovenia) **43**(2) (2019). https://doi.org/10.31449/inf.v43i2.2179
28. Wang, L., et al.: Unsupervised degradation representation learning for blind super-resolution. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10576–10585 (2021)
29. Wang, X., et al.: ESRGAN: enhanced super-resolution generative adversarial networks. In: Computer Vision - ECCV 2018 Workshops Proceedings, Part V. Lecture Notes in Computer Science, vol. 11133, pp. 63–79. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-11021-5_5
30. Wang, X., et al.: Real-ESRGAN: training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1905–1914 (2021)
31. Yeung, A., et al.: Patient motion image artifacts can be minimized and re-exposure avoided by selective removal of a sequence of basis images from cone beam computed tomography data sets: a case series. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. **129**(2), e212–e223 (2020)
32. Zhang, K., et al.: Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**, 3142–3155 (2017)
33. Zhang, K., et al.: Learning a single convolutional super-resolution network for multiple degradations. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3262–3271 (2018)
34. Zhang, K., et al.: Deep plug-and-play super-resolution for arbitrary blur kernels. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1671–1681 (2019)



3D Segmentation with Fully Trainable Gabor Kernels and Pearson's Correlation Coefficient

Ken C. L. Wong(✉) and Mehdi Moradi

IBM Research – Almaden Research Center, San Jose, CA, USA
clwong@us.ibm.com

Abstract. The convolutional layer and loss function are two fundamental components in deep learning. Because of the success of conventional deep learning kernels, the less versatile Gabor kernels become less popular despite the fact that they can provide abundant features at different frequencies, orientations, and scales with much fewer parameters. For existing loss functions for multi-class image segmentation, there is usually a tradeoff among accuracy, robustness to hyperparameters, and manual weight selections for combining different losses. Therefore, to gain the benefits of using Gabor kernels while keeping the advantage of automatic feature generation in deep learning, we propose a fully trainable Gabor-based convolutional layer where all Gabor parameters are trainable through backpropagation. Furthermore, we propose a loss function based on the Pearson's correlation coefficient, which is accurate, robust to learning rates, and does not require manual weight selections. Experiments on 43 3D brain magnetic resonance images with 19 anatomical structures show that, using the proposed loss function with a proper combination of conventional and Gabor-based kernels, we can train a network with only 1.6 million parameters to achieve an average Dice coefficient of 83%. This size is 44 times smaller than the original V-Net which has 71 million parameters. This paper demonstrates the potentials of using learnable parametric kernels in deep learning for 3D segmentation.

1 Introduction

The convolutional layer and the loss function are two fundamental components in deep learning. Because of the success of conventional deep learning kernels, i.e., the kernels with weights directly trainable through backpropagation, advancements in deep learning architectures are mainly on combining existing layers and inventing new non-convolutional layers for better performance. On the other hand, traditional parametric kernels, such as the Gabor kernel, become less popular. In fact, the versatility of conventional deep learning kernels comes with the cost of enormous numbers of network parameters proportional to the kernel size. In contrast, parametric kernels are less versatile but more compact. In this paper, we focus on the Gabor kernel as it can provide features at different frequencies, orientations, and scales, which are important for image analysis.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 53–61, 2022.
https://doi.org/10.1007/978-3-031-21014-3_6

Different frameworks have been proposed to benefit from Gabor kernels in deep learning for 2D image classification. In [10], some trainable conventional kernels were replaced by fixed Gabor kernels for more energy-efficient training. In [5], the trained conventional kernels in all convolutional layers were modulated by fixed Gabor filters to enhance the scale and orientation information. In [2], the first convolutional layer was composed of Gabor kernels with the sinusoidal frequency trainable by backpropagation. In [6], the first convolutional layer was composed of Gabor kernels, where the sinusoidal frequencies and the standard deviations were trained by the multipopulation genetic algorithm. Although the results were promising, these frameworks require manual selections of some or all Gabor parameters. This diminishes the benefits of using deep learning and manual selections can be difficult for 3D problems.

For the loss functions for multi-class image segmentation, there are mainly three categories: pixel-based [8], image-based [1,7,9], and their combinations [12]. Pixel-based losses apply the same function on each pixel and their average value is computed. A popular pixel-based loss is the categorical cross-entropy [8], which is robust to hyperparameters but may lead to suboptimal accuracy [1,12]. Image-based losses, such as the Dice loss [7], Jaccard loss [1], and Tversky loss [9], compute the losses from the prediction scores of all pixels in an image using statistical measures. The image-based losses can achieve better accuracy than the pixel-based ones, but are less robust to hyperparameters under certain situations [12]. The pixel-based and image-based losses can complement each other by weighted combinations, though deciding the optimal weights is nontrivial. Therefore, it can be beneficial if we can find a loss function that is accurate, robust to hyperparameters, and does not require manual weight selections.

To address these issues, we propose two contributions in this paper. **I)** In 3D segmentation, it is difficult to manually decide the Gabor parameters, and this can introduce unnecessary kernels while the GPU memory is precious. To gain the benefits of Gabor kernels while keeping the advantages of automatic feature generation in deep learning, we propose a Gabor-based kernel whose parameters are fully trainable through backpropagation. By modifying the formulation of a 3D Gabor kernel, we improve the versatility of the proposed Gabor-based kernel while minimizing the memory footprint for 3D segmentation. To the best of our knowledge, this is the first work of using fully trainable Gabor kernels in deep learning for 3D segmentation. Moreover, this work shows the feasibility of using only parametric kernels for spatial convolution in deep learning. **II)** We propose a loss function based on the Pearson's correlation coefficient (PCC loss) which is robust to learning rate and provides high segmentation accuracy. Different from the Dice loss which is formulated by relaxing the integral requirement of the Dice coefficient (F_1 score), the Pearson's correlation coefficient is formulated for real numbers so no approximation is required. The PCC loss also makes the full use of prediction scores from both foreground and background pixels of each label, thus is more comprehensive than the categorial cross-entropy and Dice loss. Furthermore, in contrast to the Tversky and combinatorial losses, there are no additional weights to be manually decided. Experiments on 43 3D brain

magnetic resonance images with 19 anatomical structures show that, with a proper combination of conventional and Gabor-based kernels, and the use of the PCC loss, we can train a network with only 1.6 million parameters to achieve an average Dice coefficient of 83%. This is a 44 times reduction in size compared with the original V-Net which has 71 million parameters [7].

2 Methodology

2.1 Convolutional Layer with 3D Gabor-Based Kernels

The real and imaginary parts of a 3D Gabor kernel can be represented as:

$$G_{re} = Ag(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\sigma}) \cos(2\pi f x' + \psi); \quad G_{im} = Ag(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\sigma}) \sin(2\pi f x' + \psi) \tag{1}$$

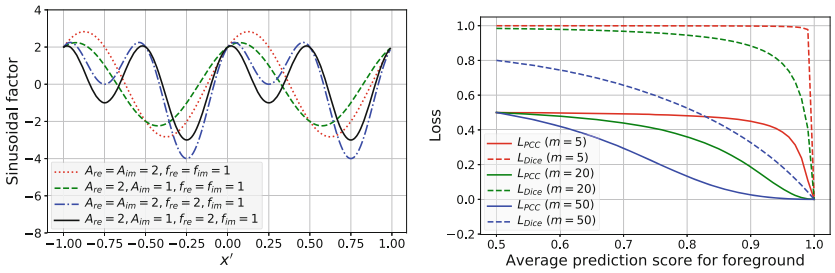
with $g(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\sigma}) = \exp\left(-0.5\left((x'/\sigma_x)^2 + (y'/\sigma_y)^2 + (z'/\sigma_z)^2\right)\right)$ the Gaussian envelope. $\boldsymbol{\theta} = (\theta_x, \theta_y, \theta_z)$ are the rotation angles about the x-, y-, and z-axis, and $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ are the standard deviations. $\mathbf{x} = (x, y, z)$ are the coordinates, and (x', y', z') are the rotated coordinates produced by $\mathbf{R} = \mathbf{R}_x(\theta_x)\mathbf{R}_y(\theta_y)\mathbf{R}_z(\theta_z)$, with $\mathbf{R}_i(\theta_i)$ the basic rotation matrix about the i-axis. A is the amplitude, f is the frequency of the sinusoidal factor, and ψ is the phase offset.

As we found that using a spherically symmetric Gaussian kernel provides similar results, we use $\sigma = \sigma_x = \sigma_y = \sigma_z$. Since $x'^2 + y'^2 + z'^2 = \|\mathbf{x}\|^2$, now only x' needs to be computed in (1) and θ_x becomes unnecessary. Although existing works either use the real and imaginary part separately or only use the real part, to increase the versatility of the Gabor-based kernel while minimizing the memory footprint for deep learning in 3D segmentation, we add the two parts together and use different A and f for the real and imaginary parts:

$$G_{DL} = g(\mathbf{x}; \sigma)\left(A_{re} \cos(2\pi f_{re} x' + \psi) + A_{im} \sin(2\pi f_{im} x' + \psi)\right) \tag{2}$$

with $g(\mathbf{x}; \sigma) = \exp\left(-0.5\left((\|\mathbf{x}\|/\sigma)^2\right)\right)$. Compared with the conventional deep learning kernel with k^3 trainable parameters, with k the kernel size, the proposed Gabor-based kernel only has eight parameters of $\{\sigma, \theta_y, \theta_z, A_{re}, A_{im}, f_{re}, f_{im}, \psi\}$. The conventional kernels can be replaced by G_{DL} in a convolutional layer, and the necessary hyperparameters of that layer remain the same: k and the number of output feature channels (c_{out}). Here k is used to form the grid of coordinates $\mathbf{x}$. Therefore, the number of trainable parameters with G_{DL} in a convolutional layer is $8 \times c_{in} \times c_{out}$, which is independent of k. As G_{DL} is differentiable, the parameters can be updated through backpropagation.

Figure 1(a) shows the characteristics of the sinusoidal factor in (2). When $f_{re} = f_{im}$, the sinusoidal factor is similar to the common sinusoidal functions even with different A_{re} and A_{im}. In contrast, more variations can be observed when f_{re} and f_{im} are different. Therefore, using different A and f for the real and imaginary parts can provide a larger variety of sinusoidal factors for better versatility.



(a) Sinusoidal factors. (b) Loss functions.

Fig. 1. (a) Sinusoidal factors with different amplitudes (A) and frequencies (f), $\psi = 0$. (b) Comparison between L_{PCC} and L_{Dice} with different cubic object size m (m^3 pixels).

2.2 Loss Function with Pearson's Correlation Coefficient

The Pearson's correlation coefficient ($r_{py} \in [-1, 1]$) measures the correlation between two variables. Given N sample pairs $\{(p_1, y_1), \ldots, (p_N, y_N)\}$, we have:

$$r_{py} = \frac{\sum_{i=1}^{N}(p_i - \bar{p})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{N}(p_i - \bar{p})^2\right)\left(\sum_{i=1}^{N}(y_i - \bar{y})^2\right)}} \tag{3}$$

where $\bar{p}$ and $\bar{y}$ are the sample means. Note that when p_i and y_i are binary, r_{py} becomes the Matthews correlation coefficient which is known to be more informative than the F_1 score (Dice coefficient) on imbalanced datasets [3].

For network training, we propose the PCC loss ($L_{PCC} \in [0, 1]$) as:

$$L_{PCC} = \mathbf{E}[1 - PCC_l]; \quad PCC_l = 0.5\left(\frac{\sum_{i=1}^{N}(p_{li} - \bar{p}_l)(y_{li} - \bar{y}_l)}{\sqrt{\left(\sum_{i=1}^{N}(p_{li} - \bar{p}_l)^2\right)\left(\sum_{i=1}^{N}(y_{li} - \bar{y}_l)^2\right) + \epsilon}} + 1\right) \tag{4}$$

where $\mathbf{E}[\bullet]$ represents the mean value with respect to semantic labels l. $p_{li} \in [0, 1]$ are the network prediction scores, $y_{li} \in \{0, 1\}$ are the ground-truth annotations, and N is the number of pixels of an image. ϵ is a small positive number (e.g., 10^{-7}) to avoid the divide-by-zero situations, which happen when all p_{li} or all y_{li} are identical (e.g., missing labels). Therefore, $L_{PCC} = 0$, 0.5, and 1 represent perfect prediction, random prediction, and total disagreement, respectively. As the means are subtracted from the samples in (3), both scores of the foreground and background pixels of each label contribute to L_{PCC}. Hence, a low L_{PCC} is achievable only if both foreground and background are well classified. This is different from the Dice loss [7]:

$$L_{Dice} = \mathbf{E}[1 - Dice_l]; \quad Dice_l = \frac{2\left(\sum_{i=1}^{N} p_{li} y_{li}\right) + \epsilon}{\left(\sum_{i=1}^{N} p_{li} + y_{li}\right) + \epsilon} \tag{5}$$

for which the background pixels do not contribute to the numerator.

Figure 1(b) shows the comparison between L_{PCC} and L_{Dice} in a simulation study. Suppose that there is a cubic image of length 100 (i.e., 100^3 pixels) with

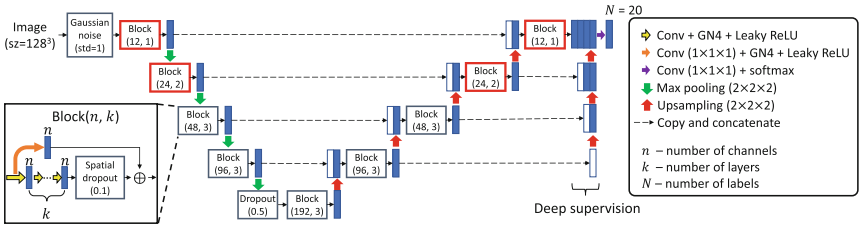


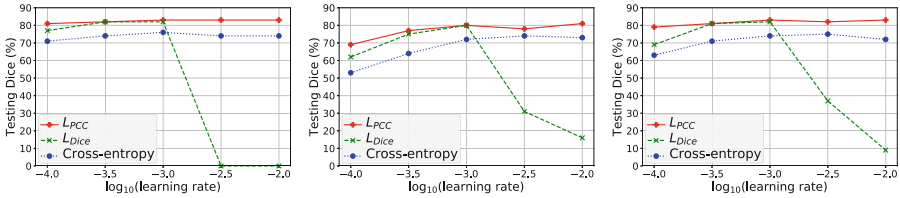
Fig. 2. Network architecture. Blue and white boxes indicate operation outputs and copied data. GN4 stands for group normalization with four groups of channels. The convolutional layers (Conv) of yellow arrows can comprise conventional kernels ($3 \times 3 \times 3$) or Gabor-based kernels ($7 \times 7 \times 7$). For the mixed-kernel models, the red blocks comprise the conventional kernels while the others comprise the Gabor-based kernels. (Color figure online)

a cubic foreground object of length m. To simulate a training process, p_{li} in the foreground are drawn from a normal distribution whose mean and standard deviation change linearly from 0.5 to 1 and from 0.5 to 0, respectively, i.e., from $\mathcal{N}(0.5, 0.5^2)$ to $\mathcal{N}(1, 0)$. For those in the background, the distribution changes linearly from $\mathcal{N}(0.5, 0.5^2)$ to $\mathcal{N}(0, 0)$. The drawn p_{li} are clipped between 0 and 1. Therefore, p_{li} change from totally random to perfect scores. In Fig. 1(b), regardless of the object size, L_{PCC} consistently starts at 0.5 and ends at 0, whereas the starting value of L_{Dice} depends on the object size. Furthermore, when the object is small ($m = 5$), the gradient of L_{Dice} is very small and suddenly changes abruptly around the prediction score of 0.99. This means that a small learning rate is required when training with L_{Dice} especially for small objects.

2.3 Network Architecture

We modify the network architecture in [12] which combines the advantages of low memory footprint from the V-Net and fast convergence from deep supervision (Fig. 2). Each block in Fig. 2 comprises the spatial convolutional layers, which can be composed of conventional or Gabor-based kernels. Spatial dropout [11] and residual connection [4] are used to reduce overfitting and enhance convergence. As the batch size is usually small for 3D segmentation because of memory requirements (e.g., one for each GPU), group normalization [13] is used instead of batch normalization for better accuracy, and four groups of channels per layer gave the best performance in our experiments.

For conventional kernels, the kernel size of $3 \times 3 \times 3$ is used as it gave good results in the experiments. For Gabor-based kernels, although the numbers of trainable parameters are independent of the kernel size, a larger kernel size is more adaptive to different frequencies and scales, and a kernel size of $7 \times 7 \times 7$ is chosen empirically. Different kernel combinations were tested (Sect. 3), including models with only conventional kernels (conventional models), with only Gabor-based kernels (Gabor-based models), and with a mix of conventional and Gabor-based kernels (mixed-kernel models). As we want the mixed-kernel models to



(a) Conventional. (b) Gabor-based. (c) Mixed-kernel.

Fig. 3. The robustness to learning rate of different loss functions with different kernel combinations. The value of each point is averaged from five experiments.

have small numbers of trainable parameters while achieving good performance, the conventional kernels are only used by the layers with fewer input and output channels, i.e., the red blocks in Fig. 2.

2.4 Training Strategy

To avoid overfitting, image augmentation with rotation (axial, ±30°), shifting (±20%), and scaling ([0.8, 1.2]) was used, and each image had an 80% chance to be transformed. The optimizer Nadam was used for fast convergence, and different learning rates and loss functions were tested in the experiments (Sect. 3). Two NVIDIA Tesla V100 GPUs with 16 GB memory were used for multi-GPU training with a batch size of two and 300 epochs.

3 Experiments

3.1 Data and Experimental Setups

A dataset of 43 3D T1-weighted MP-RAGE brain magnetic resonance images was used. The images were manually segmented by highly trained experts, and each had 19 semantic labels of brain structures. Each image was resampled to isotropic spacing, zero padded, and resized to 128 × 128 × 128.

Five dataset splits were generated by shuffling and splitting the dataset, each with 60% for training, 10% for validation, and 30% for testing. The validation sets were used to choose the best epoch in each training. Three kernel combinations, including the conventional models, Gabor-based models, and mixed-kernel models (Sect. 2.3), were tested with three loss functions of L_{PCC}, L_{Dice}, and categorical cross-entropy [12]. Each of these nine combinations was tested with five learning rates (10^{-4}, $10^{-3.5}$, 10^{-3}, $10^{-2.5}$, 10^{-2}) on the five splits. Therefore, 225 models were trained. Because of the page limit, we only compare with the basic loss functions, and their combinations are not presented.

3.2 Results and Discussion

Figure 3 shows the comparisons among the loss functions with respect to learning rates. Regardless of the kernel combinations, L_{PCC} was the most robust

Table 1. Semantic brain segmentation at learning rate 10^{-3}. (a) Semantic labels and their relative sizes on average (%). CVL represents cerebellar vermal lobules. (b) Testing Dice coefficients between prediction and ground truth averaged from five experiments (format: mean±std%). The best results are highlighted in blue.

(a) Semantic labels and their relative sizes on average (%).

1. Cerebral grey	(50.24)	2. 3rd ventricle	(0.09)	3. 4th ventricle	(0.15)	4. Brainstem	(1.46)
5. CVL I-V	(0.39)	6. CVL VI-VII	(0.19)	7. CVL VIII-X	(0.26)	8. Accumbens	(0.07)
9. Amygdala	(0.21)	10. Caudate	(0.54)	11. Cerebellar grey	(8.19)	12. Cerebellar white	(2.06)
13. Cerebral white	(31.23)	14. Hippocampus	(0.58)	15. Inf. lateral vent.	(0.09)	16. Lateral ventricle	(2.11)
17. Pallidum	(0.25)	18. Putamen	(0.73)	19. Thalamus	(1.19)		

(b) Average testing Dice coefficients (mean±std%) with respective to the ground truth.

Conventional (4.99 million parameters)																			
L_{PCC}	1.	**88±1**	2.	**80±2**	3.	**85±1**	4.	**91±0**	5.	**84±0**	6.	75±1	7.	80±1
	8.	**71±2**	9.	**77±1**	10.	**86±1**	11.	**90±0**	12.	**87±1**	13.	**90±1**	14.	**82±1**
	15.	**65±1**	16.	**91±0**	17.	**81±1**	18.	**88±1**	19.	**90±0**	**Average: 83±0**			
L_{Dice}	1.	87±1	2.	**80±2**	3.	**85±0**	4.	**91±1**	5.	83±1	6.	74±1	7.	80±1
	8.	70±1	9.	76±2	10.	85±1	11.	89±0	12.	**87±1**	13.	88±1	14.	81±1
	15.	64±3	16.	90±1	17.	**81±1**	18.	87±1	19.	89±0	**Average: 82±1**			
Gabor-based (1.53 million parameters)																			
L_{PCC}	1.	86±0	2.	78±2	3.	84±1	4.	88±1	5.	81±1	6.	71±1	7.	78±1
	8.	68±2	9.	72±2	10.	84±1	11.	87±1	12.	85±1	13.	88±0	14.	78±1
	15.	62±3	16.	89±0	17.	78±1	18.	85±2	19.	87±0	**Average: 80±0**			
L_{Dice}	1.	83±1	2.	78±1	3.	84±0	4.	88±1	5.	80±1	6.	72±1	7.	77±2
	8.	67±1	9.	71±1	10.	84±1	11.	86±0	12.	85±1	13.	87±0	14.	78±1
	15.	62±2	16.	89±0	17.	78±1	18.	85±1	19.	86±0	**Average: 80±0**			
Mixed-kernel (1.60 million parameters)																			
L_{PCC}	1.	87±1	2.	**80±2**	3.	**85±1**	4.	**91±1**	5.	83±0	6.	75±1	7.	**81±1**
	8.	70±1	9.	**77±1**	10.	85±2	11.	**90±0**	12.	**87±1**	13.	89±0	14.	**82±1**
	15.	**65±2**	16.	90±1	17.	**81±0**	18.	87±1	19.	89±0	**Average: 83±0**			
L_{Dice}	1.	86±1	2.	79±2	3.	**85±1**	4.	**91±0**	5.	83±1	6.	**76±1**	7.	80±1
	8.	70±1	9.	75±1	10.	84±2	11.	89±0	12.	**87±1**	13.	88±0	14.	81±0
	15.	64±2	16.	90±0	17.	80±2	18.	86±2	19.	88±1	**Average: 82±0**			

and accurate one among the loss functions, while the categorial cross-entropy was also robust but less accurate. L_{Dice} performed better than the categorical cross-entropy at learning rate $\leq 10^{-3}$, but its performance dropped abruptly at larger learning rates. All loss functions had their performance decreased when the learning rates $< 10^{-3}$, and the decrease of L_{Dice} was more obvious than L_{PCC}. Comparing among different kernel combinations, the conventional models performed best in general, and the mixed-kernel models outperformed the Gabor-based models. Nevertheless, if we only concentrate on L_{PCC}, the conventional and mixed-kernel models had similar performance. They also had similar performance with L_{Dice} at learning rates $10^{-3.5}$ and 10^{-3}. Moreover, the conventional models were less tolerant to L_{Dice} at larger learning rates.

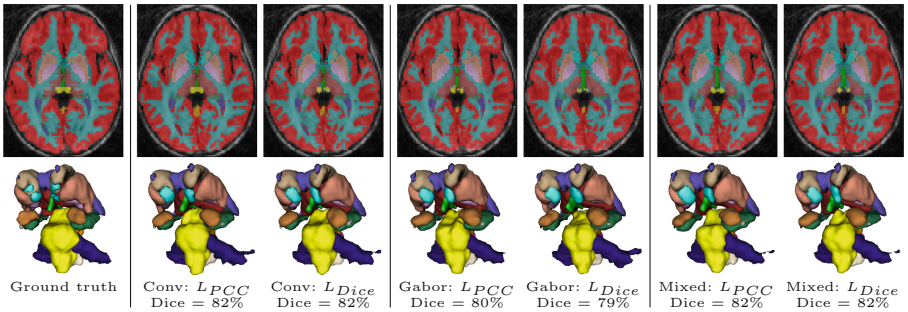


Fig. 4. Visualization of an example. Top: axial view. Bottom: 3D view with the cerebral grey, cerebral white, and cerebellar grey matters hidden for better illustration.

As all loss functions performed well at learning rate 10^{-3}, the detailed comparisons at this rate are shown in Table 1. Those of categorical cross-entropy are not shown because of their relatively low accuracy. The numbers of parameters of the conventional, Gabor-based, and mixed-kernel models were 4.99, 1.53, and 1.60 millions, respectively, thus the sizes of the conventional models were more than three times of the other models. If the kernel size of the conventional kernels changes from $3 \times 3 \times 3$ to $5 \times 5 \times 5$, i.e., the kernel size used by the V-Net [7], the numbers of parameters of the conventional and mixed-kernel models become 22.84 and 1.95 millions, respectively, more than a ten-fold difference. Table 1 also shows that the conventional and mixed-kernel models performed similarly well with less than 1% difference in Dice coefficients. The differences between using L_{PCC} and L_{Dice} were also less than 1%. Furthermore, the overall framework was very robust to network initializations as the standard deviations from five dataset splits were less than 1% on average. Note that although the Gabor-based models had the worst performance, they still had an average Dice coefficient of 80% with the least numbers of parameters.

Figure 4 shows the visualization of an example. Although the Dice coefficients of the Gabor-based models were 2% to 3% lower than the other models, their segmentations were very similar to the ground truth.

From the experimental results, we learn that L_{PCC} was more robust than L_{Dice} and more accurate than the categorical cross-entropy. The accuracies of different kernel combinations were very similar especially between the conventional and mixed-kernel models, but the mixed-kernel models used much fewer numbers of parameters. Such differences in size can be more obvious if larger kernel sizes are used. Although the Gabor-based models had the worst performance among the tested models, they still provided an average Dice coefficient of 80%. This is a good demonstration that parametric kernels can be learned through backpropagation in deep learning with decent performance.

4 Conclusion

In this paper, we propose a fully trainable Gabor-based kernel and a loss function based on the Pearson's correlation coefficient. Experimental results show that

L_{PCC} is robust to learning rate and can achieve high segmentation accuracy, and proper combinations of conventional and Gabor-based kernels can result in accurate models that are multiple times smaller than the conventional models.

References

1. Berman, M., Rannen Triki, A., Blaschko, M.B.: The Lovász-Softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4413–4421 (2018)
2. Chen, P., Li, W., Sun, L., Ning, X., Yu, L., Zhang, L.: LGCN: learnable Gabor convolution network for human gender recognition in the wild. IEICE Trans. Inf. Syst. **102**(10), 2067–2071 (2019)
3. Chicco, D.: Ten quick tips for machine learning in computational biology. BioData Mining **10**(1), 35 (2017)
4. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
5. Luan, S., Chen, C., Zhang, B., Han, J., Liu, J.: Gabor convolutional networks. IEEE Trans. Image Process. **27**(9), 4357–4366 (2018)
6. Meng, F., Wang, X., Shao, F., Wang, D., Hua, X.: Energy-efficient Gabor kernels in neural networks with genetic algorithm training method. Electronics **8**(1), 105 (2019)
7. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: IEEE International Conference on 3D Vision, pp. 565–571 (2016)
8. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
9. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: Wang, Q., Shi, Y., Suk, H.-I., Suzuki, K. (eds.) MLMI 2017. LNCS, vol. 10541, pp. 379–387. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_44
10. Sarwar, S.S., Panda, P., Roy, K.: Gabor filter assisted energy efficient fast learning convolutional neural networks. In: IEEE/ACM International Symposium on Low Power Electronics and Design, pp. 1–6 (2017)
11. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 648–656 (2015)
12. Wong, K.C.L., Moradi, M., Tang, H., Syeda-Mahmood, T.: 3D segmentation with exponential logarithmic loss for highly unbalanced object sizes. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11072, pp. 612–619. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_70
13. Wu, Y., He, K.: Group normalization. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_1



A More Design-Flexible Medical Transformer for Volumetric Image Segmentation

Xin You[1,2], Yun Gu[1,2(✉)], Junjun He[3], Hui Sun[3], and Jie Yang[1,2(✉)]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China
{geron762,jieyang}@sjtu.edu.cn
[2] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
[3] SenseTime Research, Beijing, China

Abstract. UNet-based encoder-decoder networks dominate volumetric medical image segmentation in the past several years. Many improvements focus on the design of encoders, decoders and skip connections. Due to the intrinsic property of convolutional kernels, convolution-based encoders suffer from limited receptive fields. To deal with that, recently proposed Transformer-based networks leveraging the self-attention mechanism build long-range dependency. However, they are highly reliable on pretrained weights from natural images. In our work, we find out ViT-based (Vision Transformer) models' performance will not decrease significantly without pretrained weights even if there is a limited data source. So we flexibly design a 3D medical Transformer for image segmentation and train it from scratch. Specifically, we introduce Multi-Scale Dynamic Positional Embeddings to ViT to dynamically acquire positional information of each 3D patch. Positional bias can also enrich attention diversities. Moreover, we give detailed reasons why we choose the convolution-based decoder instead of recently proposed Swin Transformer blocks after preliminary experiments on the decoder design. Finally, we propose the Context Enhancement Module to refine skipped features by merging low and high-frequency information via a combination of convolutional kernels and self-attention modules. Experiments show that our model is comparable to nnUNet on segmentation performance of Medical Segmentation Decathlon (Liver) and VerSe'20 datasets when trained from scratch.

Keywords: 3D medical segmentation · Transformer · Train from scratch

1 Introduction

Developing automatic, accurate and well-generalized medical image segmentation models has always been an essential problem in medical image analysis. With

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_7.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 62–71, 2022.
https://doi.org/10.1007/978-3-031-21014-3_7

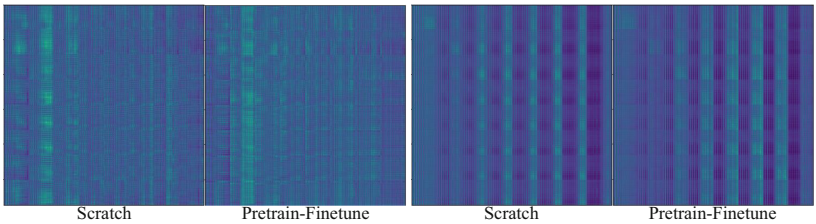


Fig. 1. Affinity maps (Inner product of query and key embeddings) of the same inputs on scratch and pretrain-finetune mode.

the development of deep learning, Convolutional Neural Networks (CNNs) have achieved great success in the field of image segmentation. Fully Convolutional Network (FCN) [12] is the first model to realize end-to-end pixel-wise semantic segmentation. UNet [17] employs a symmetric CNN encoder-decoder structure with skip connections to achieve excellent performance in many medical image segmentation tasks, which is so widely used in the field of medical image segmentation that researchers are devoted to optimizing the UNet structure.

In summary, there are three kinds of structural improvements for UNet. **Encoder:** Some works [4,8] focus on enlarging receptive fields with deeper convolutional layers and atrous convolutions. nnUNet [9] deals with local receptive fields by introducing cascaded UNets. However, the intrinsic property of convolutional kernels (Convs) still makes itself suffer from limited receptive fields. Due to the success of Transformer [21] in computer vision [6,11], such as ViT and Swin Transformer, researchers apply Transformer to medical image segmentation tasks [7,10,20] in order to better capture long-term dependency and global semantic information, which is what CNNs fail to do. Nevertheless, Transformer-based models [2,3,26] are highly dependent on pretrained weights from natural scenes, so it is inconvenient to design flexible network structures under this limitation. **Decoder:** According to recent researches [2,16], decoders are designed based on Swin Transformer blocks [11], which is different from previous CNN decoders. But those works have no direct comparisons and detailed analysis about advantages and disadvantages between two decoders. Specifically, Swin-UNet [2] and nnFormer [26] choose a symmetric Swin-based encoder-decoder design, while Swin UNETR employs a Swin encoder and a CNN decoder. **Skip connections:** For UNet, skip connections play an important role in two aspects: (a) Enrich detail information of foregrounds to compensate for information loss after downsampling and upsampling operations. (b) Backward gradient propagation makes the network converge faster, which serves as a drive force for no pretrained weights. Previous researches such as Attention UNet [14], intends to aggregate features from different channels with an attention gate. UNet++ [27] employs dense skip connections to concatenate skipped features of the same scale. BiO-Net [23] introduces bi-directional skip connections to reuse features in a recursive way. However, A recent work [15] tells that Convs behave as high-pass filters, which means skip connections purely made up of Convs may not filter out high-frequency noise. Besides, skipped features may have some important low-frequency information lost, which will affect final segmentation performance.

Table 1. Preliminary Experiments on training mode and decoder design. Swin blocks are added to decoder part bottom-to-up. (m-Dice: mean dice. m-hd_{95}: mean hd_{95})

Datasets	Pretrain		Scratch		+ 1 Swin-block		+ 2 Swin-blocks	
	m-Dice	m-hd95	m-Dice	m-hd95	m-Dice	m-hd95	m-Dice	m-hd95
Liver(MSD)	80.50	13.42	79.98	18.63	78.68	29.03	77.44	29.87
VerSe'20	84.62	4.76	83.50	4.82	82.82	5.59	82.10	5.74

The success of Transformer in natural scenes' segmentation demonstrates its potential in medical domains. However, there exist some other limitations: 1) Massive segmentation models are designed for 2D images [3,7], while medical scans are usually present as 3D types. Simply stacking slices' predictions will ignore spatial positional relations among neighboring slices. 2) Positional embeddings are of vital importance to multi-head self attention (MHSA) modules due to its permutation-invariant property, that is why absolute positional embeddings (APE) and relative positional embeddings (RPE) [11] are applied to Transformer. However, these positional encodings are fixed once training finished so that different patches of 3D scans have the same positional information.

To deal with limitations above, we redesign a 3D medical Transformer structure, corresponding to the encoder, decoder, skip connections respectively. Firstly, we have some preliminary experiments on 3D TransUNet-based network. As is shown in Table 1, we find out pretrained weights do not improve model's performance significantly. Besides, Fig. 1 indicates that affinity maps in MHSA from the last Transformer layer are similar to each other between mode scratch and pretrain-finetune, which means we can remove pretrained weights and propose a design-flexible ViT-based network. Based on this finding, we introduce Multi-Scale Dynamic Positional Embeddings (MS-DPE) to Transformer in order to dynamically acquire positional embeddings of different 3D patches. Multi-scale Convs provide a wider view for positional embeddings. Secondly, from Table 1, we can see that CNN decoders behave to be more suitable for volumetric segmentation tasks. Here in our work, we give a detailed description why we choose Convs as the basic component of the decoder instead of Swin blocks. Finally, inspired by [15], We propose the Context Enhancement Module (CEM) to effectively refine skipped features with an elaborate fuse of MHSA and Convs. Enhanced skipped features are full of low and high frequency information from foregrounds and backgrounds. Experiments demonstrate that our model is comparable to nnUNet on the segmentation performance of Liver from Medical Segmentation Decathlon(MSD) [1] and MICCAI VerSe'20 datasets [18].

2 Methodology

Overview. Our proposed model, as depicted in Fig. 2, consists of a hierarchical encoder-decoder structure. Here we employ ResNet-50 as the CNN encoder. In front of four encoder blocks, the Patch Resizing Module (PRM) is introduced to crop patches with a random ratio (0.5–2.0) to the fixed input size, then resample cropped patches into the input size. This module works due to the

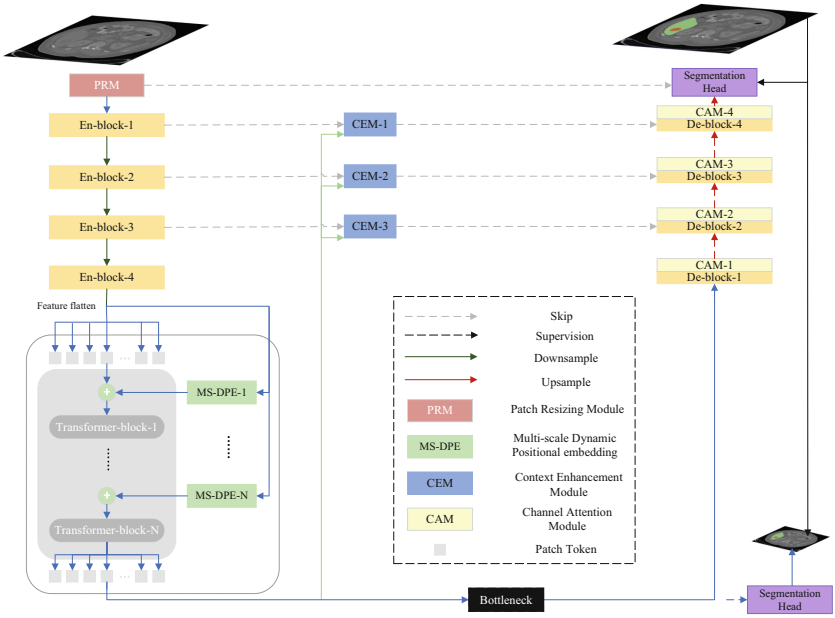


Fig. 2. Overview of Network Architecture. (a) En-block: Encoder block (b) De-block: Decoder block

fact that a larger patch size will improve segmentation performance [3]. Because the whole network is trained from scratch, we can more easily modify vanilla ViT structures. To relieve attention redundancy, here we reduce the number of Transformer layers in the ViT-B model, then introduce MS-DPE to each Transformer block. We choose Convs and residual connections as the basic component of each decoder block, with the channel attention module followed. Due to the fact that MHSA and Convs bear complementary properties, CEM is designed for skipped features with richer context. In the sections below, we will give detailed descriptions of each component mentioned above.

MS-DPE: On the ground that we can flexibly modify the structure of ViT, we put forward MS-DPE to introduce positional bias for each Transformer layer. Positional information in foreground areas is different from that in backgrounds [6]. Hence, we propose MS-DPE to dynamically generate positional embeddings as input patches vary, which is different from APE and RPE. Besides, inspired by [25], we embeds multi-scale information into tokens with CONVs for a stronger inductive bias. Detailedly, we choose 3 groups of kernel size, $1 \times 1 \times 1$, $3 \times 3 \times 3$ and $5 \times 5 \times 5$. To reduce computation cost, we use depth-wise convolutions instead. Here the introduction of convolution can increase model's generalization performance with a strong inductive bias. And multi-scale kernels are designed for global-view positional embeddings, due to limited receptive fields of flattened features generated by the CNN encoder in Fig. 2. Furthermore, MS-DPE could

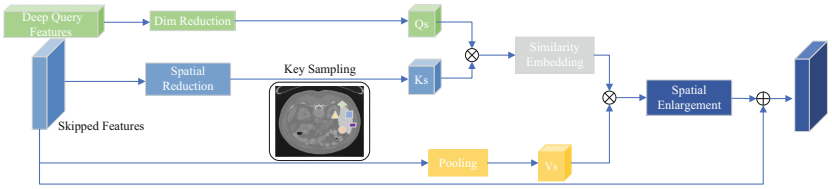


Fig. 3. Overview of Context Enhancement Module. Dim Reduction: channel dimension of deep query features are reduced to be $Q_s \in \mathbb{R}^{n \times d}$. Here n refers to token number, d represents the dimension of token vector. Spatial Reduction: spatial dimension of raw skipped features are reduced to be $K_s \in \mathbb{R}^{n \times d}$.

relieve attention redundancy [11] so that Transformer is able to generate richer global features [22] served as better queries for CEM.

Decoder Design: Here we adopt Convs and residual connections instead of Swin Transformer blocks with Patch Expanding Modules [2]. On the whole, we make this decision based on the following arguments. **1)** Swin blocks are more data-hungry for a better inductive bias [25] compared with CNNs. Besides, Shifted-Window MHSA modules bring a great challenge to the deployment of CPU devices, and its complex design makes networks more difficult to reach the global optimal point. **2)** Skipped features will enrich upsampled features with high-frequency details. Whereas, MHSA is a low-pass filter [15], so detail information may be ruined after continuous MHSA modules. **3)** For natural images' segmentation, mainstream methods will fuse multi-scale features for direct upsampling operations [4,24] due to the encoder with strong representation abilities. In medical image domains, researchers adopt the encoder-decoder structure with skip connections [17] to strengthen upsampled features. In the decoder part, three different upsampling patterns, interpolation, transposed convolution and re-permutation [11] are all equivalent to transposed convolution theoretically. Besides, Convs exploit features by gathering information from neighborhood pixels. Hence, CNN decoders are good at performing image segmentation via local information on condition that there is a well-trained encoder, that is why Swin-UNETR [19] achieves such impressive performance. The detailed decoder structure can be seen in the **Supplementary Material**.

CEM: Skipped features in encoder-decoder structures could provide more detail information to upsampled features. Skip connections purely made up of Convs may not filter out high-frequency noise due to the property of high-pass filters. Besides, skipped features may have some important low-frequency information lost, which will affect segmentation performance. Inspired by [15], MHSA and Convs bear complementary properties. Therefore, we propose CEM, which is a novel cross-attention module, to effectively refine raw skipped features with an elaborate fusion of MHSA and Convs. Enhanced skipped features can character-

Table 2. Comparison with other methods. All models are trained from scratch. (Dice1: Liver. Dice2: Liver Tumor. Dice-C: Cervical. Dice-T: Thoracic. Dice-L: Lumbar. m-D: mean dice. m-H: mean hd_{95})

Method	Param	FLOPs	Liver (MSD)				VerSe'20			
			Dice1	Dice2	m-D	m-H	Dice-C	Dice-T	Dice-L	m-D
3DUNet [5]	16.5M	803.1G	92.75	31.35	62.05	39.76	77.82	84.24	75.17	80.39
VT-UNet [16]	20.7M	123.0G	94.21	52.71	73.46	29.56	70.18	71.59	51.55	66.39
Swin UNETR [19]	72.3M	698.9G	94.92	60.89	77.90	25.28	80.29	82.68	80.61	81.51
nnUNet [9]	31.2M	534.1G	**96.35**	73.14	84.75	14.89	**88.21**	86.62	79.56	85.37
Ours	56.8M	165.7G	95.99	**74.70**	**85.35**	**8.22**	87.29	**87.63**	**82.00**	**86.18**

ize both low and high-frequency information from foregrounds and backgrounds.

$$Attention(Q_s, K_s, V_s) = Softmax(\frac{Q_s K_s^T}{\sqrt{d}})V_s \tag{1}$$

Specifically, the query Q_s comes from outputs of Transformer layers, which is good at mining key points in raw skipped features. The Key K_s is a learnable matrix via Convs for spatial reduction, and the value V_s is simply sampled with pooling operations. As is shown in Fig. 3, elements in Key Sampling relate to regions of interest. After calculating the weighted sum of V_s by Eq. 1, we need to map results into the original size, same as raw skipped features. Therefore, we can acquire enhanced skipped features by an addition of spatially enlarged results with original skipped features.

3 Experiment

3.1 Experiment Settings

Dataset. We evaluate models' performance on two public datasets, Liver from MSD [1] (train: 131 and test: 70) and MICCAI VerSe'20 [18] (train: 113, validation: 103 and test: 103). For Liver's predictions, we randomly select 20 testing cases from training data. Data splits of VerSe'20 remain unchanged. Moreover, we choose the dice similarity coefficient (DSC) and hd_{95} as quantitative metrics.

Loss Function. Our network is trained with weighted dice and cross entropy loss. As is shown in Fig. 2, total loss functions consist of two parts: supervision on final predictions $\mathcal{L}_1$ and predictions by features from bottleneck $\mathcal{L}_2$. Detailed computing process is illustrated as follows: $\mathcal{L}_i = \mathcal{L}_i^{ce} + \mathcal{L}_i^{dice}, i = 1, 2$. $\mathcal{L} = \mu\mathcal{L}_1 + \nu\mathcal{L}_2$. Because the importance of $\mathcal{L}_1$ outperforms that of $\mathcal{L}_2$, we empirically identify μ and ν as 0.7 and 0.3.

Implementation Details. We choose 4 ViT layers with 12 heads, and other parameters of ViT are the same as the ViT-Base model. We train all models using the AdamW [13] optimizer with a warm-up cosine scheduler of 1k iterations. All experiments use a batch-size of 2 per GPU (with a $128 \times 192 \times 192$

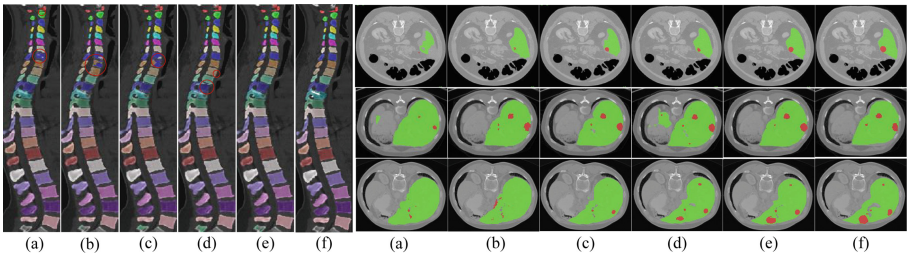


Fig. 4. Segmentation visualization of the methods above. VerSe'20: (from top to down are Cervical, Thoracic and Lumbar respectively). Liver: (Green Area: Liver. Red Area: Tumor). From left to right: (a). 3DUNet (b). VT-UNet (c). Swin UNETR (d). nnUNet (e). Ours (f). GT (Color figure online)

Table 3. Ablation Studies on liver datasets for proposed modules and decoder design.

Settings	DPE	MS-DPE	PRM	CEM	Dice1	Dice2	m-D	m-H
Baseline					95.10	64.85	79.98	18.63
Baseline +	✓				94.65	67.85	81.25	12.47
Baseline +		✓			95.58	68.04	81.81	12.07
Baseline +			✓		95.33	66.64	80.99	9.72
Baseline +				✓	**96.15**	71.09	83.62	**8.01**
Baseline +		✓	✓		95.58	70.25	82.92	8.79
Baseline +		✓	✓	✓	95.99	**74.70**	**85.35**	8.22

patch size), initial learning rate of $5e-4$, and weight decay of $1e-5$. A five-fold cross-validation strategy is used to train models with data augmentations including random flip and rotation, each fold lasting 2000 epochs. All models are implemented in PyTorch 1.7.0 and trained on 4 NVIDIA Tesla V100.

3.2 Experiment Results

Table 1 shows that the baseline model learned from scratch on Liver datasets could achieve a DSC of 79.98%, and the finetuned model only achieves a 0.52% higher DSC. On the whole, even if there is a limited data source, the ViT-based model's performance does not decrease significantly. And this phenomenon also proves that there exists a domain gap between natural and medical images.

Since pretrained weights can not bring a significant improvement to the baseline model, we train the network with proposed modules from scratch. As is illustrated by Table 2, our model's performance is comparable to nnUNet [9], with an increase of 0.60% on mean liver DSC, 1.56% on the dice of liver tumors. Visualizations of different models are shown in Fig. 4. As for VerSe'20, our model surpasses that in nnUNet by 0.81% mean DSC although nnUNet behaves better on cervical vertebrae. Visualizations in Fig. 4 show our model has a more

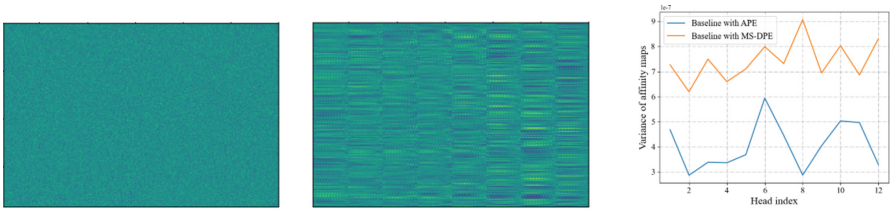


Fig. 5. (a) APE (b) MS-DPE (c) Variance of affinity maps in different heads.

consistent segmentation mask with fine boundaries. We also make a comparison with Swin-UNETR [19], which achieved SOTA performance on MSD Challenge Leaderboard. As is suggested that there exists a large gap between two models, which proves pretrained weights based on self-supervised-learning are critical to the Swin-based encoder. Besides, VT-UNet behaves inferior to 3DUNet on liver and vertebrae segmentation, which indicates CNNs bear more inductive bias and is less data-hungry compared with Swin-based networks [25].

3.3 Ablation Studies

We evaluate the effectiveness of each module via ablation studies on Liver datasets. **Effects of MS-DPE:** We choose APE as the setting of the baseline model. By adopting DPE instead of APE, performance of the model increases by 1.27% on mean DSC. After introducing multi-scale information, there exists a further 0.56% promotion on mean DSC. We visualize APE and MS-DPE from a fixed patch. As shown in Fig. 5, the visualization of MS-DPE contains richer positional information because data variance is positively related to information contents. Besides, the baseline model with MS-DPE bears more diverse attention maps according to (c) in Fig. 5. **Effects of PRM:** As indicated from Table 3, PRM could improve the baseline model by 1.01% on mean DSC. Particularly, there is a 1.79% increase on the DSC of liver tumor, proving that scaling input patches is beneficial to the segmentation of small objects. Furthermore, the model's performance can be improved by capturing sufficient contextual information. **Effects of CEM:** We employ deep features generated by Transformer as query embeddings to refine skipped features. Table 3 reveals that CEM can largely promote the segmentation accuracy of liver and liver tumor, which is a strong proof that CEM can strengthen skipped features by capturing low and high-frequency information from foregrounds and backgrounds. More visualization results are present in the **Supplementary Material**.

4 Conclusion

Based on preliminary experiments on the training mode and decoder design, we flexibly design a ViT-based 3D medical Transformer network from three components(encoder, decoder, skip connections). Firstly, we introduce MS-DPE to

dynamically acquire positional embeddings. Then we talk about why we choose Convs as the decoder instead of Swin-Transformer blocks. Finally we introduce CEM to effectively merge low and high-frequency features via a combination of Convs and MHSA. Experiments demonstrate that our proposed model is comparable to nnUNet on Liver and VerSe'20 datasets when trained from scratch.

References

1. Antonelli, M., et al.: The medical segmentation decathlon. arXiv preprint arXiv:2106.05735 (2021)
2. Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
3. Chen, J.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
6. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Gao, Y., Zhou, M., Metaxas, D.N.: UTNet: a hybrid transformer architecture for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 61–71. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_6
8. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: CCNet: criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 603–612 (2019)
9. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
10. Ji, Y., et al.: Multi-compound transformer for accurate biomedical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 326–336. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_31
11. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
14. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)

15. Park, N., Kim, S.: How do vision transformers work? In: International Conference on Learning Representations (2021)
16. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: A volumetric transformer for accurate 3D tumor segmentation. arXiv preprint arXiv:2111.13300 (2021)
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
18. Sekuboyina, A., et al.: Verse: a vertebrae labelling and segmentation benchmark. arXiv. org e-Print archive (2020)
19. Tang, Y., et al.: Self-supervised pre-training of swin transformers for 3D medical image analysis. arXiv preprint arXiv:2111.14791 (2021)
20. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: gated axial-attention for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 36–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_4
21. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
22. Wu, Y., et al.: D-former: a U-shaped dilated transformer for 3D medical image segmentation. arXiv preprint arXiv:2201.00462 (2022)
23. Xiang, T., Zhang, C., Liu, D., Song, Y., Huang, H., Cai, W.: BiO-Net: learning recurrent bi-directional connections for encoder-decoder architecture. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 74–84. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_8
24. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11209, pp. 432–448. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01228-1_26
25. Xu, Y., Zhang, Q., Zhang, J., Tao, D.: ViTAE: vision transformer advanced by exploring intrinsic inductive bias. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
26. Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnFormer: interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)
27. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1



Dcor-VLDet: A Vertebra Landmark Detection Network for Scoliosis Assessment with Dual Coordinate System

Han Zhang(✉), Tony C. W. Mok, and Albert C. S. Chung

Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{hzhangcp,cwmokab,achung}@cse.ust.hk

Abstract. Spinal diseases are common and difficult to cure, which causes much suffering. Accurate diagnosis and assessment of these diseases can considerably improve cure rates and the quality of life for patients. The spinal disease assessment relies primarily on accurate vertebra landmark detection, such as scoliosis assessment. However, existing approaches do not adequately exploit the relationships between vertebrae and analyze the global spine structure, meaning scarcity annotations are underutilized. In addition, the practical design of ground-truth is also deficient in model learning due to the suboptimal coordinate system. Therefore, we propose a unified end-to-end vertebra landmark detection network called Dcor-VLDet, contributing to the scoliosis assessment task. This network takes the positional information from within and between vertebrae into account. At the same time, through fusing the advantages of both Cartesian and polar coordinate systems, the symmetric mean absolute percentage error (SMAPE) value can be reduced significantly in scoliosis assessment. The experimental results demonstrate that our proposed method is superior in measuring Cobb angle and detecting landmarks on low-contrast X-ray images.

Keywords: Vertebra landmark detection · Scoliosis assessment · Convolutional neural network

1 Introduction

Spine diseases are diverse and occur at all ages. Scoliosis is a kind of spine disease with three-dimensional spine deformity, that includes coronal, sagittal, and axial sequence abnormalities. Scoliosis can impact a child's or adolescent's growth and development, distort the body, affect cardiopulmonary function, and even involve the spinal cord, that in extreme cases results in paralysis [10]. Scoliosis is a prevalent condition that puts adolescents and youngsters in danger. Clinically, anterior-posterior (AP) X-ray images have been used for diagnosis. However, manual annotation and quantifying the scoliosis severity are too prohibitive and daunting in terms of time and resources.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 72–80, 2022.
https://doi.org/10.1007/978-3-031-21014-3_8

To overcome these limitations, medical image analysis researchers have paid much attention to the use of computer-aided diagnosis (CAD) systems for effective diagnosis. BoostNet [11] introduces a BoostLayer and a spinal structured multi-output regression layer for dealing with adolescent idiopathic scoliosis assessment. Yi et al. [12] presents a vertebra-focused landmark detection approach based on keypoint detection for the scoliosis assessment problem. By localizing the vertebra centers, the network discriminates between individual vertebrae. After collecting the vertebrae, several convolutional layers are utilized to regress the four corner landmarks of each vertebra. Based on [1,12] describes a multi-task technique that guarantees semantic masks and keypoints are consistent, as well as a keypoint Transformer to determine the spine structure. The MPF-net, proposed in [13], combines a vertebra detection branch with a landmark prediction branch to provide a restricted region for landmark prediction. The information between adjacent vertebrae may then be extracted using a proposal correlation module and a feature fusion module.

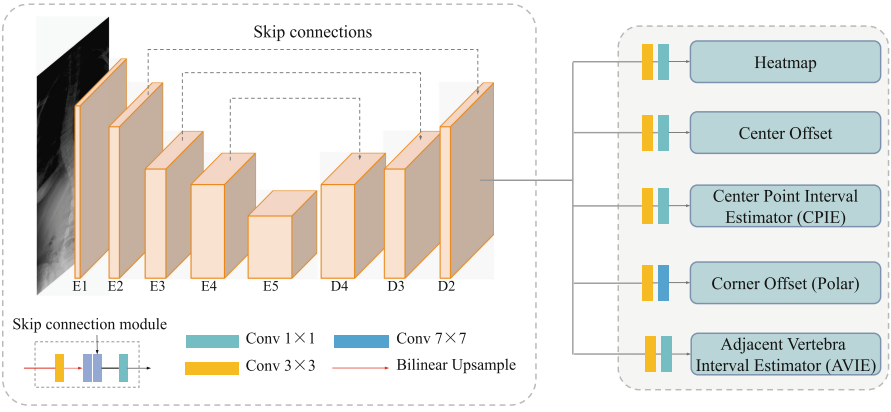


Fig. 1. The architecture of the Dcor-VLDet network. The encoder and decoder are represented by E and D, respectively.

Although the above research works have achieved specific results, they do not study the structure and characteristics between vertebrae or integrate domain knowledge into the design of the network. In the anterior-posterior spinal x-ray images, the upper part has less interference from the background, while the lower part might be overlaid with other tissue, making the localization challenging. In this case, only considering local information is not sufficient for vertebra localization, especially in the lower part. Based on these considerations, exploiting explicit and implicit information from the existing annotations according to the domain knowledge becomes exceptionally significant. However, none of the approaches above fully use the positional correlation between vertebrae, which may be derived from the existing annotations without extra annotations. As a result, we present the Dcor-VLDet, an end-to-end landmark detection network,

that considers the correlation between vertebrae and fuses dual coordinate systems. Dcor-VLDet alleviates the problem by adding relative center point offsets and adjacent vertebra interval offsets as additional supervision information. We further make ground-truth more effective by combining multiple coordinate systems based on the properties of the sub-tasks.

To summarize, the contributions of our work are listed as follows:

- Introducing a center point interval estimator (CPIE), which can provide inter-vertebra supervision information for predicting the center points of vertebrae. It improves center point localization, especially for the vertebrae with significant background interference, reducing error accumulation caused by erroneous center points.
- Introducing an adjacent vertebra interval estimator (AVIE) as an auxiliary task, which can better exploit the implicit information of the ground-truth and provide more supervision information for the corner point localization.
- Proposing a dual coordinate system learning scheme, in which both Cartesian and polar coordinate systems are adopted to represent the ground-truth of center points and corner offsets, respectively, which can better preserve the strengths of both coordinate systems in the sub-tasks.

2 Method

2.1 Vertebra Landmark Detector Framework

In this paper, a novel dual coordinate system-based model, namely Dual Coordinate System Vertebra Landmark Detector (Dcor-VLDet), is proposed. The intrinsic method used by most of the recent vertebra landmark detectors, Cartesian coordinates modeling, is replaced by dual coordinate system modeling. As shown in Fig. 1, following the decoder, there are five pathways for further landmark decoding. One of them is the heatmap pathway, which is used to estimate the center points of the vertebrae. The other two pathways, the center offset and the corner offset pathway, are used to regress the Cartesian coordinates of the center points and the polar coordinates of four corner points. Another two pathways, CPIE and AVIE have been proposed and incorporated into the Dcor-VLDet model for more accurate center point and corner point regression. There are five loss terms as shown in Eq. 1: the heatmap loss (hm), the center offset regression loss (center), the corner offset regression loss (corner), the CPIE loss and the AVIE loss, the hyper-parameter α_1 to α_5 are 1. They are optimized jointly. As the backbone of the whole detector, a modified pre-trained ResNet34 [2] is selected as the encoder architecture of Dcor-VLDet to extract semantic features from the X-ray images. Then, skip connections are used to combine deep features with shallow features.

$$\mathcal{L} = \alpha_1 \mathcal{L}_{hm} + \alpha_2 \mathcal{L}_{center} + \alpha_3 \mathcal{L}_{corner} + \alpha_4 \mathcal{L}_{CPIE} + \alpha_5 \mathcal{L}_{AVIE}. \quad (1)$$

In the inference phase, we estimate and identify the predicted center point as the maximum response of the predicted heatmap as presented in [5,12]. To

optimize the model parameters, we employ the same focal loss as described in [12] with the same parameter setting. The objective of the center offset pathway is to reduce the quantization error caused by downsampled inputs and bilinear interpolation. The points, which are mapped from the position (h, w) in the original input image, are represented by the position $(\lfloor h/k \rfloor, \lfloor w/k \rfloor)$ of the downsampled feature map, where k is the downsampling factor [12]. The definition of the center offset is $(h/k - \lfloor h/k \rfloor, w/k - \lfloor w/k \rfloor)$, and the parameters are optimized using the L1 loss.

2.2 Dual Coordinate System

We trace the four corner points based on the positions of the center points through the corner offsets. Corner offsets are described as vectors that begin at the center and point to the corners of the vertebrae. As each vertebra presents essentially symmetrical features, the distances from the center point to the four corner points in each vertebra are similar. Therefore, the corner points represented by polar coordinates with the center points as the origins can better focus on the corner points' distances and angles from the center points. However, this information would not be straightforward enough for the network to learn when the Cartesian coordinate system is used. Instead, when employing the polar coordinate system, these minor differences within distances and angles become more noticeable for the network, making the corner point localization process more precise. On the contrary, the localization of the center points performs better under the Cartesian coordinate system since the localization of center points is less dependent on a predefined origin. As a result, the Cartesian coordinate system can be used to better identify the positions of the center points in the image. Owing to the discrepancies between center points and corner points, multiple coordinate systems are adopted to ensure the positioning of both is more accurate. The joint utilization of dual coordinate systems can help develop the network's learning capabilities to a greater extent while avoiding the decline of the learning performance caused by the irrationality of ground-truth coordinates.

In a Cartesian coordinate system, four corner points of each vertebra are frequently expressed as (x_1^k, y_1^k), (x_2^k, y_2^k), (x_3^k, y_3^k), and (x_4^k, y_4^k) as illustrated in Fig. 2(d), where k specifies the index of the vertebra. To describe them in polar coordinates, we first define the vertebra center point as the polar coordinate system's pole point. Then, the positive direction of the polar axis is defined by the horizontal-right direction, while the polar angle in radians is specified as counterclockwise. The four corners can be represented in the polar coordinate system as (r_1^k, θ_1^k), (r_2^k, θ_2^k), (r_3^k, θ_3^k), and (r_4^k, θ_4^k), and serve as the ground-truth of the corner offsets for training. The corner offset pathway is trained using the L1 loss. Since the process of evaluating the performance of vertebra landmark detection is currently only carried out in Cartesian coordinates, we need to transform the corner point in polar coordinates to a Cartesian one during the inference phase. First, heatmap and center offset outputs are used to extract the locations of the center point (x_{ct}^k, y_{ct}^k) in Cartesian coordinates. Then, using the transformation formula, Eq. 2 [14], the four corner points in the form of $[(x_1^k,$

y_1^k), (x_2^k, y_2^k), (x_3^k, y_3^k), and (x_4^k, y_4^k)] may be obtained, where n indicates the indices of the corner points in each vertebra.

$$x_n^k = x_{ct}^k + r_n^k * \cos(\theta_n^k), \quad y_n^k = y_{ct}^k + r_n^k * \sin(\theta_n^k). \tag{2}$$

2.3 Center Point Interval Estimator and Adjacent Vertebra Interval Estimator

From the spinal X-ray image in Fig. 2(a), it is obvious that the extent of background interference is different in the upper and lower parts of the spine. According to the extent of background interference severity, we divide 17 vertebrae across the spine into two groups, namely the upper ten vertebrae and the lower seven vertebrae, as shown in Fig. 2(a). The background of the upper part is mainly the lung, while the background of the lower part is the thoracic and abdominal cavity, making the lower part vertebrae more challenging to locate. The mean detection error (MDE) in Table 1, which computes in pixels, can provide more quantitative results for the localization of 17 center points and 68 landmarks. Using the baseline method, the MDE for the lower part's center points is around 50% larger than that in the upper part, resulting in an increase of around 10 pixels in the overall MDE for the 17 vertebrae. Simultaneously, the center point localization errors will accumulate to the localization of corner points, resulting in the calculation error of the Cobb Angle for the scoliosis assessment. Therefore, to enhance the localization accuracy of center points, we propose the CPIE. The center point intervals are illustrated in Fig. 2(a). By introducing CPIE as an auxiliary pathway following the decoder, the correlation between the center points in adjacent vertebrae can be completely explored and play an active role in the training process through back-propagation process. Furthermore, multi-task learning acts as a regularizer by introducing an inductive bias [8], and prevents the training process from becoming biased towards any particular task. As a result, the risk of overfitting is also reduced [8].

As demonstrated in Fig. 2(b), the predicted corner points using the baseline method are obviously not precise enough, and the deviation is mostly due to the fact that they are not situated in the border corners of the vertebrae but rather in the interior of the vertebrae. An intuitive reason for this is that the boundary between two adjacent vertebrae is not sufficiently differentiated. Therefore, we need to strengthen the network's learning capacity on the boundary information between two adjacent vertebrae. An ideal approach is to give the model as much supervision information on adjacent vertebra boundaries as feasible to learn without incurring additional labeling costs. Therefore, we propose AVIE as one pathway following the decoder of the backbone. As illustrated in Fig. 2(d), we set the corner points for each vertebra in a specific order. As demonstrated in Fig. 2(d), the offset is from the corner points (x_3^{k-1}, y_3^{k-1}), (x_4^{k-1}, y_4^{k-1}) of the upper one vertebra to the corner point (x_1^k, y_1^k), (x_2^k, y_2^k) of the lower one vertebra respectively. The parameters in both estimators are fine-tuned using the L1 loss throughout the training phase.

Table 1. The MDE values of the vertebra landmarks and center points on the AASCE dataset. CT, CT_{upper} and CT_{lower} denote the center points of the entire spine, the upper part and lower part of the spine, respectively. LM, LM_{upper} and LM_{lower} denote the landmarks of the entire spine, the upper part and lower part of the spine, respectively. The MDE values are measured in pixels.

Method	Auxiliary estimator		Corner point obtain			CT↓	CT_{upper}↓	CT_{lower} ↓	LM↓	LM_{upper}↓	LM_{lower}↓
	CPIE	AVIE	Relative	Absolute	Polar						
Baseline [12]			✓			60.33	49.81	75.36	63.19	52.09	79.07
Ours	✓		✓			55.18	44.89	69.89	58.68	47.61	74.50
Ours		✓	✓			55.45	44.95	70.45	59.13	47.72	75.43
Ours	✓	✓	✓			**48.54**	**39.73**	**61.13**	**52.06**	**42.53**	**65.67**

3 Experiment

3.1 Dataset and Implementation Details

We evaluate our proposed method on the AASCE MICCAI 2019 challenge dataset [6]. The AASCE dataset contains 609 spinal X-ray images in the anterior-posterior plane. Each image contains 17 vertebrae from the thoracic and lumbar spine. Four corner landmarks are provided for each vertebra. The Cobb angle is estimated with AASCE's algorithm. We resize the X-ray images to 1024×512 as the network's inputs. On this X-ray image dataset, we employ the same data split as reported in [12], which contains 60% samples for training, 20% samples for validation, and 20% samples for testing.

All experiments have been conducted with PyTorch 1.10 on a PC with a single NVIDIA 2080 GPU. The model was optimized using the Adam optimizer [3]. The network's other weights are initialized by a standard Gaussian distribution. The initial learning rate is set to 1.25×10^{-4}. All models are trained for 100 epochs. The backbone ResNet34 [2] is pre-trained with the ImageNet dataset [4].

3.2 Evaluation Metrics

To evaluate the method's performance quantitatively, we measure the MDE values between the detected landmarks and the ground-truth landmarks as the evaluation metric. The MDE value mentioned above can be computed through:

$$MDE = \frac{1}{N} \sum_{j=1}^{N} \left\| \text{out}_j - \text{gt}_j \right\|_2, \tag{3}$$

where N denotes the total number of landmarks in the test dataset. out_j and gt_j denote the output landmarks and the ground-truth landmarks, respectively. On NVIDIA 2080, the frames per second (FPS) are recorded for comparing the efficiency of the models.

For the scoliosis assessment, the Cobb angle is the gold standard [9]. SMAPE has been used for Cobb angle measurement as shown in Eq. 4. The proximal thoracic (PT), main thoracic (MT), and thoracolumbar (TL)'s Cobb angles are also

Table 2. The vertebra landmark detection and scoliosis assessment performance of different methods and the ablation study on the AASCE dataset can be seen below. All MDE values are given as a mean ± std and are measured in pixels. SMAPE is abbreviated to SP.

Method	Auxiliary estimator		Corner point obtain			SP↓	SP_{PT}↓	SP_{MT}↓	SP_{TL}↓	MDE↓	FPS↑
	CPIE	AVIE	Relative	Absolute	Polar						
Baseline [12]			✓			9.82	5.68	15.77	22.15	63.19±67.36	20.54
Ours	✓		✓			8.82	5.27	15.22	20.27	58.68±67.56	**27.72**
Ours		✓	✓			9.10	5.60	15.66	21.88	59.13±65.29	26.67
Ours	✓	✓	✓			8.61	4.98	14.59	21.48	**52.06**±64.27	26.56
Ours				✓		9.61	6.30	14.67	21.67	62.61±58.10	27.52
Ours					✓	8.36	5.04	16.21	20.21	55.61±63.49	12.66
Ours	✓	✓		✓		11.53	8.13	18.77	21.82	66.19±68.47	16.04
Ours	✓	✓			✓	**7.53**	**4.65**	**13.74**	**18.81**	57.61±68.62	26.28

measured [7] and reported as SMAPE_{PT}, SMAPE_{MT} and SMAPE_{TL}, respectively. In Eq. 4, p signifies the Cobb angles of PT, MT, and TL, q means the q-th image, while M denotes the total number of testing images. The *out* and *gt* stand for the estimated and ground-truth Cobb angles, respectively.

$$\text{SMAPE} = \frac{1}{M}\sum_{q=1}^{M}\frac{\sum_{p=1}^{3}(|\text{out}_{qp} - \text{gt}_{qp}|)}{\sum_{p=1}^{3}(\text{out}_{qp} + \text{gt}_{qp})}. \tag{4}$$

3.3 Experimental Results

To evaluate the center point and corner point localization capabilities of both the upper and lower parts of the spine applying the proposed interval estimators. Some quantitative results have been listed in Table 1. As shown in Table 1, these two interval estimators perform well in capturing the center points, especially in the lower part as opposed to the upper part of the spine. It only reduces MDE by 10.08 pixels for the upper part whereas for the lower part it reduces it by 14.23 pixels. The localization accuracy for corner points also follows the same trend. The SMAPE value decreases from 9.82 to 8.61 using the proposed two estimators as shown in Table 2.

Table 2 demonstrates the experimental results comparing the baseline approach to our methods and the results of the ablation study. Unlike the results in Table 1, which mainly compare the performance of different regions of the spine, Table 2 contains the experimental results for all 68 landmarks along the whole spine using different models. We compare different evaluation metric results by using 1) relative offsets from the center points to the corner points in the Cartesian coordinate system (relative), 2) the absolute Cartesian coordinates in the whole image (absolute), and 3) the relative offset from the center point in the polar coordinate system (polar). The baseline results are shown in the first row of Table 2, where only the relative Cartesian coordinates are used for the corner points, the same as described in [12]. We report the results of six evaluation

metrics produced with and without CPIE and AVIE. We apply the same data augmentation techniques given in [12] to all of the baseline and our approaches, including random cropping, expanding, contrast, and brightness distortion.

In Table 2, the lowest overall MDE value is 52.06 pixels, which is excessively high. However, this is primarily caused by the misclassification of the vertebrae. A typical example is shown in Fig. 2(b). There are over 17 vertebrae in this X-ray image. However, only 17 vertebrae are annotated. These incomplete annotations will cause errors in capturing the order of vertebrae and computing MDE between vertebrae, which do not correspond to one another. Therefore, some cases with substantial MDE errors are generated. Although our strategy can make center points more accurate and reduce cases with high MDE to a certain extent, some such cases cannot be avoided. We noticed that the dual coordinate system could contribute to the proposed Dcor-VLDet with the SMAPE declined by 1.46. Some qualitative results on landmark detection and corner offset regression are shown in Fig. 2.

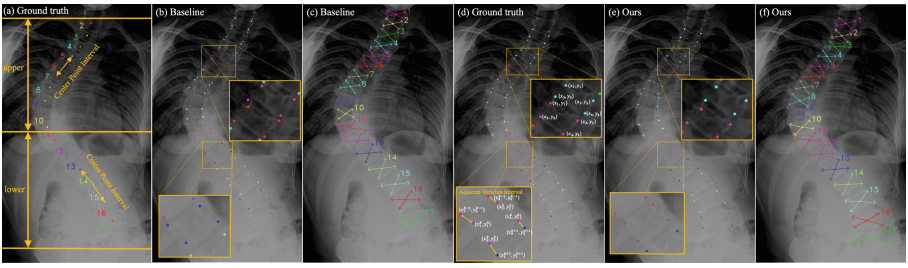


Fig. 2. Qualitative results using the Dcor-VLDet model. (a) and (d) are ground-truth. (b) and (c) show the vertebra landmark detection results of the baseline method [12] and (e) and (f) show the vertebra landmark detection results of our method.

4 Conclusion

Through the proposed Dcor-VLDet network, we have made contributions to the scoliosis assessment task. It can preserve the benefits of the Cartesian coordinate system to aid in the localization of the center points, while fully exploit the polar system's advantages in locating corner points. These two interval estimators suggested can provide additional supervision information for a more accurate center point and corner point localization procedure. When the approaches mentioned above are incorporated into a unified model, we discover that our novel network outperforms the baseline landmark detection network in landmark detection and scoliosis assessment through comprehensive experiments on the AASCE dataset.

References

1. Guo, Y., Li, Y., Zhou, X., He, W.: A keypoint transformer to discover spine structure for cobb angle estimation. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2021)

2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
3. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst. **25**, 1097–1105 (2012)
5. Law, H., Deng, J.: CornerNet: detecting objects as paired keypoints. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 765–781. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_45
6. Li, S., Wang: Accurate automated spinal curvature estimation MICCAI 2019 (2019). https://aasce19.github.io/
7. O'Brien, M., Group, S.D.S.: Radiographic Measurement Manual. Medtronic Sofamor Danek USA (2008). https://www.oref.org/docs/default-source/default-document-library/sdsg-radiographic-measuremnt-manual.pdf?sfvrsn=2&sfvrsn=2
8. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
9. Scholten, P., Veldhuizen, A.: Analysis of cobb angle measurements in scoliosis. Clin. Biomech. **2**(1), 7–13 (1987)
10. Weinstein, S.L., Dolan, L.A., Cheng, J.C., Danielsson, A., Morcuende, J.A.: Adolescent idiopathic scoliosis. The Lancet **371**(9623), 1527–1537 (2008)
11. Wu, H., Bailey, C., Rasoulinejad, P., Li, S.: Automatic landmark estimation for adolescent idiopathic scoliosis assessment Using BoostNet. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 127–135. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_15
12. Yi, J., Wu, P., Huang, Q., Qu, H., Metaxas, D.N.: Vertebra-focused landmark detection for scoliosis assessment. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 736–740. IEEE (2020)
13. Zhang, K., Xu, N., Guo, C., Wu, J.: MPF-Net: an effective framework for automated cobb angle estimation. Med. Image Anal. **75**, 102277 (2022)
14. Zhou, L., Wei, H., Li, H., Zhao, W., Zhang, Y.: Objects detection for remote sensing images based on polar coordinates. arxiv 2020. arXiv preprint arXiv:2001.02988



Plug-and-Play Shape Refinement Framework for Multi-site and Lifespan Brain Skull Stripping

Yunxiang Li[1,2], Ruilong Dan[2], Shuai Wang[3], Yifan Cao[2], Xiangde Luo[4], Chenghao Tan[2], Gangyong Jia[2], Huiyu Zhou[5], You Zhang[1], Yaqi Wang[2,6(✉)], and Li Wang[7(✉)]

[1] Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, USA
[2] Hangzhou Dianzi University, Hangzhou, China
[3] School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China
[4] University of Electronic Science and Technology of China, Chengdu, China
[5] School of Computing and Mathematical Sciences, University of Leicester, Leicester, UK
[6] Communication University of Zhejiang, Hangzhou, China
wangyaqi@cuz.edu.cn
[7] Developing Brain Computing Lab, Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, USA
li_wang@med.unc.edu

Abstract. Skull stripping is a crucial prerequisite step in the analysis of brain magnetic resonance images (MRI). Although many excellent works or tools have been proposed, they suffer from low generalization capability. For instance, the model trained on a dataset with specific imaging parameters cannot be well applied to other datasets with different imaging parameters. Especially, for the lifespan datasets, the model trained on an adult dataset is not applicable to an infant dataset due to the large domain difference. To address this issue, numerous methods have been proposed, where domain adaptation based on feature alignment is the most common. Unfortunately, this method has some inherent shortcomings, which need to be retrained for each new domain and requires concurrent access to the input images of both domains. In this paper, we design a plug-and-play shape refinement (PSR) framework for multi-site and lifespan skull stripping. To deal with the domain shift between multi-site lifespan datasets, we take advantage of the brain shape prior, which is invariant to imaging parameters and ages. Experiments demonstrate that our framework can outperform the state-of-the-art methods on multi-site lifespan datasets.

Keywords: Skull stripping · Transformer · Domain adaptation · Shape dictionary · Lifespan brain

Y. Li and R. Dan—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_9.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 81–90, 2022.
https://doi.org/10.1007/978-3-031-21014-3_9

1 Introduction

Skull stripping, the separation of brain tissue from non-brain tissue, is a critical preprocessing step for the characterization of brain MRI. Plenty of skull stripping tools have been proposed, e.g., morphology-based method: Brain Surface Extractor (BSE) [1] and surface-based method: Brain Extraction Tool (BET) [2]. Compared with traditional skull stripping tools, deep learning has recently been proven more suitable for skull stripping, where 3D U-Net is the most popular backbone [3,4]. Based on it, 3D Residual U-Net, 3D Attention UNet, and TransBTS are proposed to get better segmentation performance [5–7]. In addition, some specific methods have been designed, e.g., Zhong et al. proposed a domain-invariant knowledge-guided attention network for brain skull stripping [8]. Zhang et al. proposed a flattened residual network for infant MRI skull stripping [9]. However, the high performance of deep learning-based methods requires that the training and testing datasets share a similar data distribution, which is hardly met due to a variety of device manufacturers, magnetic field strength, and acquisition protocols. Moreover, there is also a substantial data distribution difference across lifespan, e.g., the adult and infant brain MRI in Fig. 1, where the infant's brain is undergoing myelination and maturation.

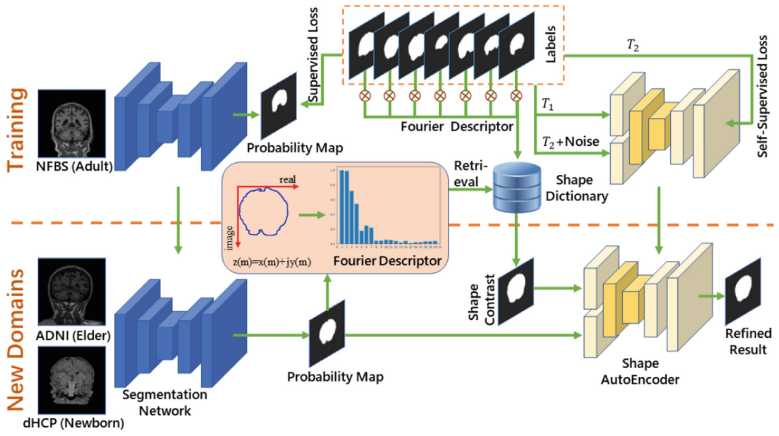


Fig. 1. A schematic overview of our proposed plug-and-play shape refinement framework.

Many efforts have been devoted to addressing the domain shift problem. Among them, the most widely used method is domain adaptation to align the latent feature distributions of the two domains [10,11]. Unfortunately, there are some inherent limitations, including the need for retraining for each new domain and concurrent access to the input images of both domains. It is well known that medical images are difficult to share due to privacy. Compared with the original image, manual labels contain much less privacy information and are

easier to share. Therefore, we propose a plug-and-play shape refinement (PSR) framework in this work. The main contributions of our method are three-fold: **1)** A novel plug-and-play segmentation result refinement framework is designed. **2)** A shape dictionary based on the Fourier Descriptors [12] is proposed to fully utilize the anatomical prior knowledge of the brain shape. **3)** To better model the overall shape information, we designed a shape AutoEncoder (SAE) based on Shuffle Transformer.

2 Method

2.1 Overall Architecture

The overall framework of PSR is illustrated in Fig. 1. Specifically, we can arbitrarily take a model trained in the source domain as the segmentation network. Fourier Descriptors [12] of the source labels are computed to build a shape dictionary. Then, the new domain image is directly input into the model, and Fourier Descriptors of the segmentation results are computed. Through it, we can retrieve a label with the closest shape from the shape dictionary. Finally, the segmentation results, together with the retrieved labels, serve as inputs into SAE to further refine the segmentation results.

2.2 Shape Dictionary

To make full use of the anatomy prior knowledge of brain shape, we calculate the corresponding Fourier Descriptors for each subject according to the source labels and store them in the dictionary. Assuming source labels $L^s = \{l_i^s\}_{i=1}^{|L^s|}$, the constructing process of shape dictionary D^s is defined as Eq. (1)

$$D^s = \{d_i^s \mid d_i^s = F(l_i^s)\}_{i=1}^{|L^s|} \tag{1}$$

where F is the Fourier Descriptor for a quantitative representation of closed shapes independent of their starting point, scale, location, and rotation. The whole process of computing Fourier Descriptors consists of three steps: (1) Establish a coordinate system in the upper left corner of the boundary, and the coordinate axis is tangent to the boundary. (2) Take the two axes as real and imaginary numbers respectively, and the coordinates of points (x_m, y_m) on the boundary are expressed in complex numbers $z(m) = x_m + jy_m$. (3) The discrete Fourier transform (DFT) is applied to the above coordinates to obtain the Fourier Descriptor of the boundary shape, which is defined as $Z(k)$ in Eq. (2).

$$Z(k) = \frac{1}{N} \sum_{m=0}^{N-1} z(m) e^{-j2\pi mk/N}, k = 0, 1, 2, ..., N-1 \tag{2}$$

where N is the amount of the boundary points. Specifically, we choose 10 low-frequency coefficients as the final Fourier Descriptors, which is sufficient to achieve the required accuracy for retrieving [13].

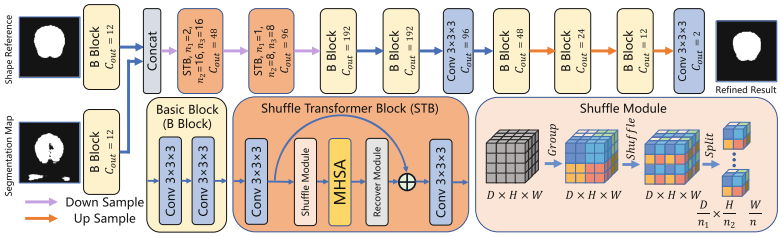


Fig. 2. A schematic overview of our Shape AutoEncoder. The skip connection of the U-shaped structure is not shown for simplicity.

2.3 Shape AutoEncoder (SAE)

The Shape AutoEncoder is designed to further refine the segmentation results based on both segmentation maps from the segmentation network and the shape reference from the shape dictionary, with the detailed structure shown in Fig. 2. To better extract global information such as overall shape, Transformer-based methods are good candidates [14–16]. Based on 2D Shuffle Transformer [17], we design a 3D Shuffle Transformer to capture global capability.

Shuffle Transformer: Our shuffle Transformer mixes the voxel features regularly, then evenly divides them into non-overlapping groups. Specifically, Shuffle Module takes the feature X as input and outputs shuffled blocks X^b; thus it can be formulated through Eq. (3).

$$X^b = \{x_i^b \mid x_i^b = Split(Shuffle(Group(X, i))), x_i^b \in \mathbb{R}^{\frac{D}{n_1} \times \frac{H}{n_2} \times \frac{W}{n_3}}\}_{i=1}^{(n_1 \times n_2 \times n_3)} \tag{3}$$

where the details of Group, Shuffle, and Split are depicted in Fig. 2. After these operations we obtain a total number of $(n_1 \times n_2 \times n_3)$ shuffled blocks x^b, each with a size of $\frac{D}{n_1} \times \frac{H}{n_2} \times \frac{W}{n_3}$. Subsequently, each group is computed by multi-head self-attention. Since our shuffle operation keeps the relative position of each element in the shuffled blocks as the same as the original feature block, our position encoding method is based on relative-distance-aware position encoding [18,19]. The query-key-value (QKV) attention [20] in each small block x^b can be computed by Eq. (4).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}} + B)V \tag{4}$$

where Q, K, and V stand the query, key, and value matrices of dimension d_k respectively, and B denotes the position biases matrix.

Self-supervised Training of Shape AutoEncoder: As for the training process of SAE, it is based on self-supervision, meaning the supervisory signals are

generated by the input, and we use the shared source labels as the training images of SAE. Each image is processed by two kinds of random spatial transformations (T_1 and T_2). Specifically, our random spatial transformations contain random scaling, random rotation, and random clipping. Assume we have images $Y = \{y_i\}_{i=1}^{\eta}$ where η denotes the number of images, thus transformed images are generated via Eq. (5).

$$\begin{aligned} Y^{T_1} &= \{y_i^{T_1} \mid y_i^{T_1} = T_1(y_i)\}_{i=1}^{\eta} \\ Y^{T_2} &= \{y_i^{T_2} \mid y_i^{T_2} = T_2(y_i)\}_{i=1}^{\eta} \end{aligned} \tag{5}$$

To mimic the unreliable segmentation results caused by domain shift, random noises RN are added to T_2 through placing random false positive and false negative stain to generate $Y^{T_2 \circ RN}$ formulated by Eq. (6). The random noise is controlled by two variables: the amount and size of the noise. The hyperparameters of the random noise are discussed in the supplementary material.

$$Y^{T_2 \circ RN} = RN(Y^{T_2}) \tag{6}$$

Let $SAE(y_i^{T1}, y_i^{T_2 \circ RN}; \theta)$ be the Shape AutoEncoder, thus refined outputs $\hat{Y}$ are defined via Eq. (7).

$$\hat{Y} = SAE(Y^{T_1}, Y^{T_2 \circ RN}; \theta) \tag{7}$$

$\hat{Y}$ and θ represent the prediction output and the trainable parameters of SAE respectively, and the self-supervised loss function of the SAE is defined by Eq. (8).

$$\begin{aligned} \mathcal{L}_{SAE}(\hat{Y}, Y^{T_2}) = -\frac{1}{\eta} \sum_{i=1}^{\eta} (T_2(y_i) \times ln(SAE(T_1(y_i), RN(T_2(y_i)); \theta)) \\ + (1 - T_2(y_i)) \times (1 - ln(SAE(T_1(y_i), RN(T_2(y_i)); \theta)))) \end{aligned} \tag{8}$$

In this way, it can not only combine segmentation results and shape reference but also denoise and refine the unreliable segmentation results automatically through learning shape prior knowledge from the labels.

3 Implementation and Experiments

3.1 Datasets and Evaluation Metrics

We evaluated the proposed method on the publicly available dataset, where the source domain is from Neurofeedback Skull-stripped (NFBS) [21], and the new (target) domains are from Alzheimer's Disease Neuroimaging Initiative (ADNI) [22] and Developing Human Connectome Project (dHCP) [23]. Note that subjects from NFDS are young adults from 21 to 45 years old and ADNI are older adults from 55 and 90 years old, and dHCP are newborns. After resampling and padding, the size of the individual scan is $256 \times 256 \times 256$ and each voxel size

is $1 \times 1 \times 1$ mm^3. We selected 25 subjects from NFBS with manual labels as the training dataset and each 10 subjects from ADNI and dHCP as the testing dataset, and 3-fold cross-validation is used. It is worth noting that there is no available publicly manual label of dHCP, and thus we only compare the results by visual inspection. Due to the limited GPU memory, a sub-volume of size $64 \times 64 \times 64$ is used as the first stage segmentation network input. In order to better capture the overall shape, the input size of the Shape AutoEncoder is set to $8 \times 256 \times 256$, and the middle layer slice with the size of 256×256 is used to compute the Fourier Descriptors and retrieve the most similar shape from the dictionary. For all experiments described below, the Average Surface Distance (ASD), the Dice Coefficients (DICE), the sensitivity (SEN) and specificity (SPE) are chosen for evaluation metrics.

3.2 Implementation Details

All our experiments were based on the PyTorch framework and carried out on 4 Nvidia RTX 2080Ti GPUs. We trained our network from scratch for a total of 10000 iterators, and the parameters were updated by the Adam algorithm (momentum = 0.97, weight decay = 5×10^{-4}). We adopt a batch size of 4 and set the learning rate as 2×10^{-4}. Notedly, if there is no additional statement, our PSR is combined with 3D U-Net by default in this paper. The discussion of the hyperparameters is on the supplementary material.

Table 1. Comparison with the state-of-the-art methods on ADNI.

Method	ASD (mm)	DICE (%)	SPE (%)	SEN (%)
3D U-Net [3]	11.57 ± 9.83	88.30 ± 3.53	99.51 ± 0.68	82.69 ± 3.10
CycleGAN [24]	18.96 ± 7.88	86.27 ± 2.85	98.72 ± 0.51	85.19 ± 2.58
EMNet [25]	8.67 ± 9.86	91.49 ± 2.39	99.37 ± 0.49	**89.32 ± 2.10**
Tent [26]	7.19 ± 6.24	90.52 ± 1.70	**99.64 ± 0.28**	85.60 ± 2.18
PSR (3D U-Net)	**4.63 ± 0.98**	**91.57 ± 1.38**	99.55 ± 0.15	88.09 ± 2.67

3.3 Experimental Results on Cross-site Dataset

Our results are presented in Table 1, and our method is combined with 3D U-Net, namely PSR (3D U-Net). We can observe that our method enhances the performance of 3D U-Net a lot. Moreover, our method outperforms the state-of-the-art domain adaptation method, i.e., EMNet. Interestingly, the ASD achieved by other methods was unexpectedly far higher than ours. This may be due to the fact that the training data from the source domain do not have the shoulder part, but the testing data in the new domain do have the shoulder, which is illustrated in Fig. 3. Our method is capable of identifying the segmentation result of the shoulder as unreliable results and excluding it from the brain tissues. We can also notice that the output of CycleGAN mistakenly identifies many non-brain

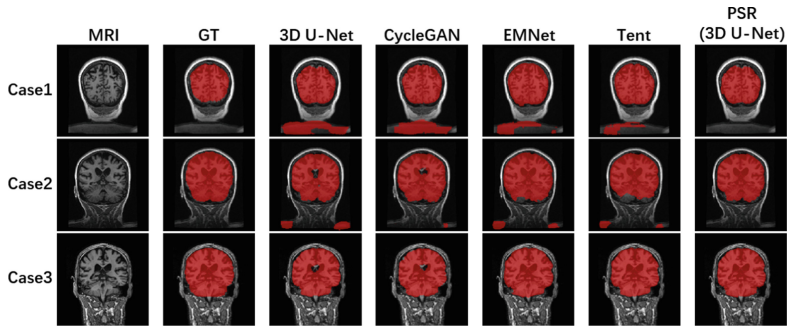


Fig. 3. Visualization of segmentation results on ADNI.

regions as the brain. Consequently, its performance is unexpectedly poorer than those obtained from the source model without domain adaptation.

In order to further test the potential of our PSR, we also combined it with other popular networks. As shown in Table 2, the segmentation performances of all the networks have been enhanced a lot by combining with our PSR.

Table 2. Comparison with the state-of-the-art methods on ADNI.

Method	ASD (mm)	DICE (%)
3D Residual U-Net	13.73 ± 9.69	88.99 ± 4.19
PSR (3D Residual U-Net)	**4.42 ± 1.61**	**91.34 ± 1.61**
3D Attention U-Net	10.13 ± 7.03	89.29 ± 2.17
PSR (3D Attention U-Net)	**4.06 ± 0.76**	**91.22 ± 0.99**
TransBTS	12.31 ± 12.26	88.80 ± 4.67
PSR (TransBTS)	**6.01 ± 2.64**	**90.87 ± 1.65**
3D U-Net	11.57 ± 9.83	88.30 ± 3.53
PSR (3D U-Net)	**4.63 ± 0.98**	**91.57 ± 1.38**

3.4 Visualization Results on Newborns

We qualitatively compare the segmentation outputs of the proposed PSR and the 3D U-Net on the newborn. Although the infant shares a similar shape and basic structure with the adult, the infant's brain develops rapidly throughout the first year of life, resulting in huge appearance differences from the adult. As shown in Fig. 4, 3D U-Net cannot achieve accurate brain tissues with fuzzy brain boundaries, while our method is able to generate smooth and reasonable segmentations.

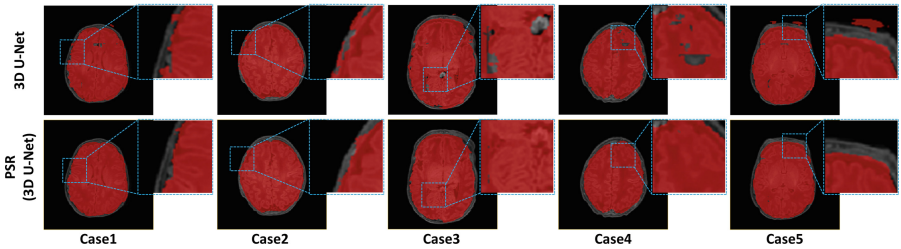


Fig. 4. Qualitative comparison of segmentation results on newborn brain MRI from dHCP.

3.5 Ablation Study

In order to verify the effectiveness of the main components of our proposed PSR, we conducted the following ablation study: a) implement our PSR without Shape AutoEncoder, referred to as PSR w/o SAE. b) implement our PSR without shape dictionary, that is, only the unrefined segmentation results are input to Shape AutoEncoder, referred to as PSR w/o SD. c) implement our PSR without shuffle Transformer, that is, replace the shuffle Transformer block with the basic convolution block, denoted as PSR w/o ST. Quantitative comparison results of the three variants along with the PSR are illustrated in Table 3. The proposed PSR achieves improved performance, especially in terms of important ASD and DICE metrics.

Table 3. Ablation study of our method.

Method	ASD (mm)	DICE (%)	SPE (%)	SEN (%)
PSR w/o SAE	11.57 ± 9.83	88.30 ± 3.53	99.51 ± 0.68	82.69 ± 3.10
PSR w/o SD	6.86 ± 3.53	89.73 ± 1.42	99.41 ± 0.13	86.04 ± 2.21
PSR w/o ST	6.39 ± 4.24	90.62 ± 1.40	**99.76 ± 0.13**	84.79 ± 2.66
PSR	**4.63 ± 0.98**	**91.57 ± 1.38**	99.55 ± 0.15	**88.09 ± .2.67**

4 Conclusion

In summary, we presented a plug-and-play shape refinement framework and successfully applied it to the skull stripping task on multi-site lifespan datasets. Our method consists of a shape dictionary and a Shape AutoEncoder. With the assistance of the shape dictionary, the Shape AutoEncoder makes full use of the anatomical prior knowledge to refine the unreliable segmentation results in the new domain. Experimental results demonstrated that the proposed method can enhance the performance of most networks and achieves better performance than the state-of-the-art unsupervised domain adaptation methods, and the proposed Shape AutoEncoder can further enhance traditional skull stripping tools. Theoretically, our method can also be widely applied to other organ segmentation.

Acknowledgements. This work was supported in part by National Institutes of Health (Grant No. R01CA240808 and R01CA258987), National Natural Science Foundation of China (Grant No. U20A20386), Shandong Provincial Natural Science Foundation (Grant No. 2022HWYQ-041)

References

1. Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M.: Magnetic resonance image tissue classification using a partial volume model. NeuroImage **13**(5), 856–876 (2001)
2. Smith, S.M.: Fast robust automated brain extraction. Human Brain Map. **17**(3), 143–155 (2002)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
4. Luo, X., et al.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 318–329. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_30
5. Islam, M., Vibashan, V.S., Jose, V.J.M., Wijethilake, N., Utkarsh, U., Ren, H.: Brain tumor segmentation and survival prediction using 3D attention UNet. In: Crimi, A., Bakas, S. (eds.) BrainLes 2019. LNCS, vol. 11992, pp. 262–272. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-46640-4_25
6. Yu, W., Fang, B., Liu, Y., Gao, M., Zheng, S., Wang, Y.: Liver vessels segmentation based on 3D residual U-Net. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 250–254. IEEE (2019)
7. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: multimodal brain tumor segmentation using transformer. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 109–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_11
8. Zhong, T., et al.: DIKA-Nets: domain-invariant knowledge-guided attention networks for brain skull stripping of early developing macaques. NeuroImage **227**, 117649 (2021)
9. Zhang, Q., Wang, L., Zong, X., Lin, W., Li, G., Shen, D.: Frnet: flattened residual network for infant MRI skull stripping. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 999–1002. IEEE (2019)
10. Li, Y., et al.: Dispensed transformer network for unsupervised domain adaptation. arXiv preprint arXiv:2110.14944 (2021)
11. Dou, Q., et al.: PnP-AdaNet: plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. IEEE Access **7**, 99065–99076 (2019)
12. Nixon, M.S., Aguado, A.S.: Chapter 7 - object description. In: Nixon, M.S., Aguado, A.S. (eds.) Feature Extraction & Image Processing for Computer Vision (Third Edition), pp. 343–397. Academic Press, Oxford (2012)
13. Dalitz, C., Brandt, C., Goebbels, S., Kolanus, D.: Fourier descriptors for broken shapes. EURASIP J. Adv. Signal Process. **2013**(1), 1–11 (2013)

14. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
15. Li, Y., et al.: AGMB-transformer: anatomy-guided multi-branch transformer network for automated evaluation of root canal therapy. IEEE J. Biomed. Health Inform. **PP**(99), 1 (2021)
16. Li, Y., et al.: GT U-Net: a U-net like group transformer network for tooth root segmentation. In: Lian, C., Cao, X., Rekik, I., Xu, X., Yan, P. (eds.) MLMI 2021. LNCS, vol. 12966, pp. 386–395. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87589-3_40
17. Huang, Z., Ben, Y., Luo, G., Cheng, P., Yu, G., Fu, B.: Shuffle transformer: rethinking spatial shuffle for vision transformer. arXiv preprint arXiv:2106.03650 (2021)
18. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
19. Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V.: Attention augmented convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019
20. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
21. Eskildsen, S.F., et al.: Beast: brain extraction based on nonlocal segmentation technique. NeuroImage **59**(3), 2362–2373 (2012)
22. Jack Jr, C.R., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging Off. J. Int. Soc. Magnetic Reson. Med. **27**(4), 685–691 (2008)
23. Makropoulos, A., et al.: The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. Neuroimage **173**, 88–112 (2018)
24. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
25. Sun, Y., et al.: Multi-site infant brain segmentation algorithms: the ISEG-2019 challenge. IEEE Trans. Med. Imaging **40**(5), 1363–1376 (2021)
26. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: TENT: fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021)



A Coarse-to-Fine Network for Craniopharyngioma Segmentation

Yijie Yu[1], Lei Zhang[1(✉)], Xin Shu[1], Zizhou Wang[1], Chaoyue Chen[2], and Jianguo Xu[2]

[1] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China
leizhang@scu.edu.cn

[2] Department of Neurosurgery, West China Hospital, Sichuan University, Chengdu, China

Abstract. Craniopharyngioma (CP) is one of the most common intracranial tumors located in the sellar region and its surroundings, which often leads to visual acuity, visual field disorders, and pituitary hypothalamus dysfunction. Segmentation of CP is an essential prerequisite in the diagnosis, screening, and treatment. Also, It's a challenging task due to the indistinguishable borders, the small tumor size, and high diversity in size, shape, and texture. In this work, a novel automatic coarse-to-fine CP segmentation network is proposed, consisting of two stages: the coarse segmentation stage and the refinement stage. During the first stage, the Coarse Segmentation Guided Module (CSGM) is proposed to generate rough segmentation results and exclude the interference of background regions. During the refinement stage, the Local Feature Aggregation (LFA) module is proposed to solve the boundary ambiguity problem. It can encode the fine-grained information and adaptively explore the dependencies between a local spatial neighborhood. To validate the effectiveness of our model, a realistic CP dataset was constructed and a 4.26% dice score promotion is achieved compared to the baseline.

Keywords: Craniopharyngioma segmentation · Coarse-to-fine · Coarse segmentation guided module · Local feature aggregation module

1 Introduction

Craniopharyngioma (CP) is a rare benign tumor originating from the remnants of Rathke's pouch, with a bimodal age distribution of 5–14 years and 50–75 years [1]. Because of the anatomical closeness of CP to critical structures, including the pituitary, hypothalamus, and visual apparatus [2,3], some long-term survivors may suffer from visual and hypothalamic dysfunction and resulting in a lower overall quality of life. In clinical practice, CP segmentation based on Magnetic Resonance Imaging (MRI) [4] is fundamental to preoperative diagnosis and treatment. So it's highly desirable to develop an automatic CP segmentation approach to assist physicians in making a confident and correct diagnosis.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 91–100, 2022.
https://doi.org/10.1007/978-3-031-21014-3_10

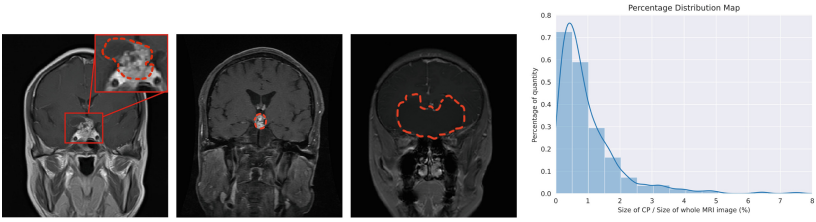


Fig. 1. Illustration of MRI images of CP, which is outlined with a red dashed line. (a) The cavernous sinus has been partly invaded by CP, making the borders indistinguishable. (b) CP with relatively small size. (c) CP with relatively large size. (d) The percentage distribution map of $\frac{\text{size of CP}}{\text{size of MRI}}$. (Color figure online)

So far, with the rapid development of deep convolutional neural networks (DCNNs) [5], some remarkable achievements have been made in the field of medical image processing. However, it's still a challenging task due to some unique characteristics of MRI images of CP: (1) CP has a strong tendency to densely adhere to and invade surrounding structures, making the borders indistinguishable (as shown in Fig. 1a). (2) The tumor areas usually comprise only a tiny percentage of the whole image, causing the severe regional imbalance between foreground and background. As shown in Fig. 1b and Fig. 1d, the tumor areas account for no more than 1% in more than 70% MRI images. (3) The shape, size, and texture of CPs can vary greatly in different MRI slices (as shown in Fig. 1b and Fig. 1c), making the models prone to CPs with large-size and underperform for small-size tumors.

To solve the first challenge, many architectures embedded with attention mechanisms are proposed, such as, Inf-Net [6], GSCNN [7], Attention UNet [8], ET-Net [9], CPFNet [10]. These models can automatically learn to suppress irrelevant representations while highlighting salient useful features. Specifically, Inf-Net [6] added the edge constraint guidance to reverse attention to explicitly enhance the boundary identification. GSCNN [7] was the first to process shape into a separate parallel stream and the Gated Convolutional Layer (GCL) is proposed for the information fusion between the two streams. However, these methods usually calculated the attention weights based on the whole feature maps, leading to a dramatic increase in computational complexity and efficiency. For the latter two problems, some methods have been developed in a coarse-to-fine framework. They often utilized two cascade networks: a lighter, faster one for coarse segmentation and a more detail-oriented one for refinement based on previous results [11–13]. At the same time, with the popularity of target detection, the coarse segmentation stage can be replaced by target detection [14,15]. For example, Ma et al. [15] tried to generate a bounding box using the Target Detection Module (TDM) during the coarse segmentation stage, which was then fed into two elaborate models according to the size of targets to achieve fine processing. But these methods have two major drawbacks: Firstly, models

with two networks often bring a cumbersome training process and redundant use of computational resources. Secondly, the error of the coarse segmentation result will be accumulated in the next stage.

In this paper, to overcome the above-mentioned challenges, a novel CP segmentation network based on UNet [16] is proposed and developed in a coarse-to-fine framework. Our architecture consists of two stages: the coarse segmentation stage and the refinement stage. During the first stage, the Coarse Segmentation Guided Module (CSGM) is proposed to generate rough segmentation results with a high recall rate and reduce the interference of irrelevant background regions. During the refinement stage, to emphasize the detailed information passed from low-level feature maps, the Local Feature Aggregation (LFA) module is proposed, adaptively exploring the dependencies between a local spatial neighborhood. Overall, the contributions of this work are: (1) To solve the boundary ambiguity problem, the LFA is proposed to efficiently select more effective detailed information (boundary, shape, etc.) from the low level while significantly reducing the computational effort. (2) To address the problem of small tumor areas and high variations in tumor size, a novel CSGM is proposed to roughly localize the tumor and remove the irrelevant background regions, allowing the architecture to achieve the coarse-to-fine framework with only one network while minimizing the accumulation of errors during the process. (3) A coarse-to-fine end-to-end network is provided for CP segmentation based on CSGM and LFA and achieves the best performance on our CP dataset.

2 Method

2.1 Overview of the Network Architecture

The architecture of the proposed method is illustrated in Fig. 2, which consists of two stages: the coarse segmentation stage and the refinement stage. The architecture employs UNet [16] as backbones, embedding the CSGM and LFA. Moreover, the non-local attention module [17] is utilized to expand the perception field and catch long-range dependencies. Next, CSGM and LFA will be introduced in detail.

2.2 Coarse Segmentation Guided Module

To address the challenges of small tumor areas and high variations in tumor size, we develop a novel CSGM (as shown in Fig. 2) to utilize the coarse segmentation results obtained from the shallower network to filter out irrelevant information passed to the network in the subsequent training process.

In the first CSGM, given a feature map f^1 obtained from the decoder block, we can get a coarse segmentation result p^1 by the following process:

$$s^1 = \sigma(\mathcal{C}(\mathcal{U}(f^1))) \tag{1}$$

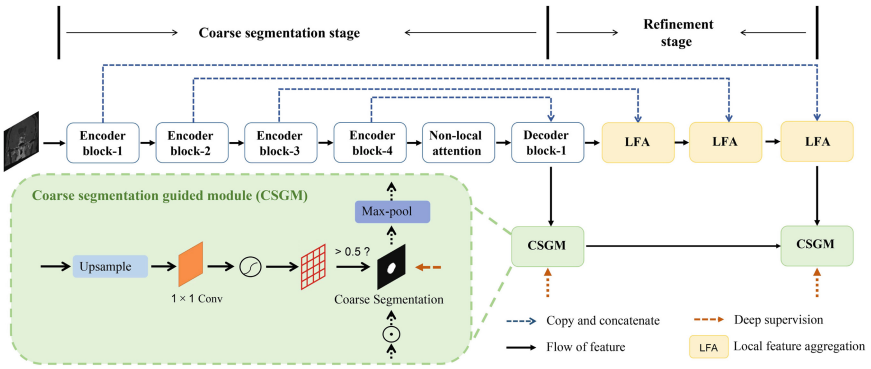


Fig. 2. Illustration of the proposed architecture and coarse segmentation guided module (CSGM).

$$p^1 = \begin{cases} 0, if & 0 < s^1 < 0.5 \\ 1, if & 0.5 \leq s^1 < 1 \end{cases} \tag{2}$$

where $\mathcal{U}(\cdot)$ denotes an up-sampling of f^1 to the size of original image, $\mathcal{C}(\cdot)$ denotes the convolutional operation with the kernel size of 1×1 and $\sigma(\cdot)$ is the Sigmoid function which is used to generate an activated score map s^1 with the values between 0 and 1. And then the coarse segmentation result p^1 can be obtained through Eq. 2. Note that $p^1 = 1/0$, which means the coarse segmentation result can be considered as being cropped from the image. To minimize the errors, Max-pooling and Tversky loss (introduced in detail in the following sub-section) are used to expand the rough segmentation areas. The final attention gate p^1_{dilate} can be obtained by:

$$p^1_{dilate} = Maxpool(p^1) \tag{3}$$

Next, p^1_{dilate} is treated as a spatial attention map and passed to the next CSGM module. In the second CSGM, the activation score map s^2 can be obtained by Eq. 1 in the same way as above, so we can get the s^2_{refine} as below:

$$s^2_{refine} = p^1_{dilate} \odot s^2 \tag{4}$$

where $\odot$ is element-wise multiplication. Then, the new refined activation score map s^2_{refine} is compared with the ground truth for supervision.

2.3 Local Feature Aggregation

Considering that most of the irrelevant regions have been excluded, constraints on the aggregation scope to a spatial local neighborhood are crucial for finer-level feature learning and reducing the computational complexity. Motivated by this, the Local Feature Aggregation (LFA) module is proposed. As shown in Fig. 3, letting f_l and f_h denote the feature maps from low-level and high-level,

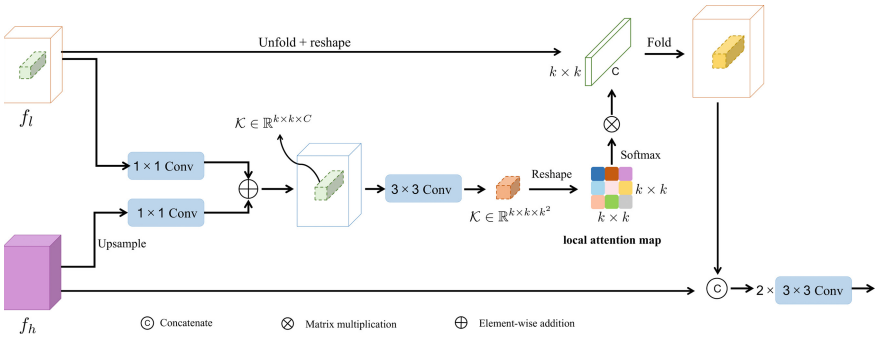


Fig. 3. Illustration of Local Feature Aggregation (LFA) Module.

where $f_l \in \mathbb{R}^{C \times H \times W}$, and C, H, W denote the channel, height and width of f_l. For each spatial location (i, j) in any feature map, we assume it's affected by a local window of size $k \times k$ centered at (i, j). So the local feature attention map $K \in \mathbb{R}^{H \times W \times k^2 \times k^2}$ can be obtained by the following procedure:

$$K = Softmax(Conv_f(\varphi(f_l) + \theta(\mathcal{U}(f_h)))) \tag{5}$$

where $\varphi(\cdot)$ and $\theta(\cdot)$ represent 1×1 convolutions with weights $\mathcal{W}_\varphi$ and $\mathcal{W}_\theta$, respectively, $\mathcal{U}(\cdot)$ indicate the upsample operation, $Conv_f(\cdot)$ is the 3×3 convolution and $softmax(\cdot)$ denotes the softmax function applied in every local window.

Given the attention map K, we multiply it with f_l to generate the refined feature map. Specifically, f_l is first unfolded and reshaped to $f_l{}'$, where $f_l{}' \in \mathbb{R}^{H \times W \times k^2 \times C}$, and then we can get the final refined low-level feature maps $f_l^{refined}$ by:

$$f_l^{refined} = K \otimes f_l^{'} \tag{6}$$

where $\otimes$ represents matrix multiplication operation. At last, all the refined sliding local blocks $f_l^{refined}$ are combined into a large containing feature map, which is then concatenated with the high-level feature and followed by two consecutive 3×3 convolution operations to get the final output.

2.4 Loss Function

In this work, deep supervision [18] is employed to every CSGM. To maximize the recall rate of the first CSGM output, Tversky loss [19] is introduced. The definition is as follows:

$$L_t = 1 - Tl \tag{7}$$

$$Tl = \frac{\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i g_i + \alpha \sum_{i=1}^{N} (1 - p_i) g_i + \beta \sum_{i=1}^{N} p_i (1 - g_i) + \epsilon} \tag{8}$$

where N is the pixel number of one slice, p_i represents the predicted probabilities of pixel i and g_i represents the binary ground truth of pixel i, ϵ represents for a hyperparameter to prevent division by zero. Hyperparameters α and β can be employed to weigh false positive (FP) and false-negative (FN) flexibly. FN can be weighted more heavily to obtain a result with higher recall. In addition, the effective binary cross-entropy loss L_{bce} is used for each output. So the total loss for each branch output can be computed as:

$$L = \mu_1 L_t + \mu_2 L_{bce} \tag{9}$$

where μ_1, μ_2 are balance terms. Given that two CSGMs are adopted in the whole network, the final loss function of our architecture can be defined as:

$$L_{total} = L_1 + L_2 \tag{10}$$

3 Experiments

3.1 Datasets and Implementation Details

So far, no public dataset for CP segmentation is available. To evaluate the proposed method, a dataset of CP MRI scans was constructed by West China Hospital, including 228 scans from 228 unique patients, which are randomly divided into three parts: 180 scans for training, 20 for validation, and 28 scans for testing. All of the scans are manually labeled by experienced neurosurgeons using 3D Slicer. Before training, the voxels' size of all the images are resampled to $1 \times 1 \times 1$, normalized to $[0, 1]$ by Min-max normalization, and resized to 256×256. When training, Adam [20] optimizer with the initial learning rate of 10^{-4} is adopted and the cosine annealing learning rate is employed where the maximum number of iterations is set to 200. The α and β in Eq. 8 are set to 0.1, 0.9 in the first CSGM module and 0.5, 0.5 in the second CSGM. The μ_1, μ_2 in Eq. 9 are set to 1, 1. The local window size k in LFA modules is set to 3, 3, and 5, respectively. The pre-train weights of the network come from ImageNet.

3.2 Results and Discussion

Compare with State-of-the-Art Segmentation Methods. The quantitative results of the experiments are represented in Table. 1. The Dice (DSC) score, Intersection of Union (IoU), Sensitivity (Sen), Precision (Pre), and 95HD are used to evaluate the segmentation results. We compared the model with several traditional medical image segmentation methods (UNet, UNet++, ResUNet++, DeepLabV3) and attention-based methods (AG-Net, CA-Net, GLFRNet, CPFNet, TransUNet) to validate the effectiveness of the proposed architecture. Compared to the baseline (UNet), our architecture outperformed the results by 4.26% for Dice, 7.41% for IoU, 1.5% for Precision, and 6.80% for Sensitivity. As observed, our method achieves the best performance in terms of DSC, IoU, Pre, $95HD$ compared to other methods. The higher $Dice$ and lower

Table 1. The comparison result of the proposed method with other SOTA methods.

Method	*DSC*	*IoU*	*Pre*	*Sen*	95HD	*#Param.*
UNet [16]	0.8151	0.6879	0.8749	0.7630	8.6154	**17.27M**
UNet++ [21]	0.8405	0.7249	0.8729	0.8104	9.7606	36.63M
ResUNet++ [22]	0.7991	0.6654	0.7445	**0.8623**	8.9096	35.46M
DeepLabV3 [23]	0.8043	0.6727	0.8328	0.7777	8.3253	59.34M
Attention UNet [24]	0.8408	0.7253	0.8786	0.8061	7.3852	34.87M
CA-Net [25]	0.8200	0.6950	0.8654	0.7792	8.7613	44.42M
CPFNet [10]	0.8298	0.7091	0.8591	0.8024	7.4638	43.27M
TransUNet [26]	0.7715	0.6280	0.8245	0.7248	8.7093	105.32M
GLFRNet [27]	0.8468	0.7343	0.8748	0.8206	7.7352	30.30M
Ours	**0.8498**	**0.7389**	**0.8878**	0.8149	**7.0290**	17.91M

$95HD$ indicate that our model works better for both boundary and body. It is worth noting that our model greatly reduces the number of parameters while improving the results. For instance, although GLFRNet achieves a comparable dice score (84.68% vs 84.98%), its number of parameters is almost twice than that of ours (30.30M vs 17.91M). The visualization of segmentation results is shown in Fig. 4. In the last row, the corona slice is incomplete, typically seen when the patient's head is non-rigidly aligned during the MRI scan process. Most of the models fail to predict the result under the situation, but the employment of CSGM in our architecture makes the network more sensitive to the tumor and its surrounding area, thus successfully localizing the tumor.

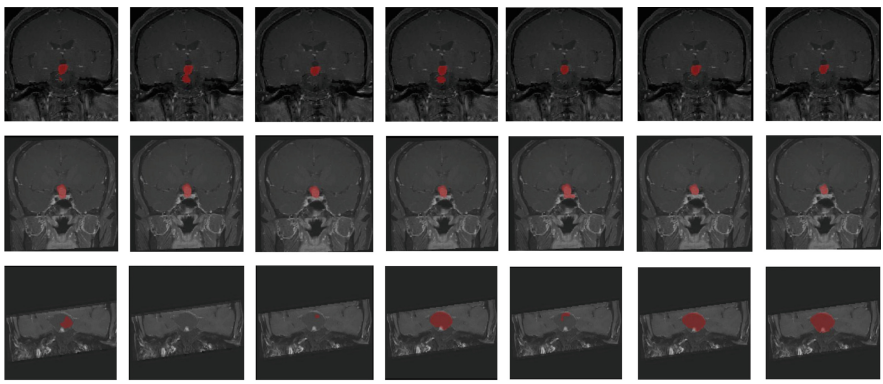


Fig. 4. Visualization of segmentation results of ours and other SOTA methods. Tumor areas are colored in red. (Color figure online)

Table 2. Ablation study on CSGM and LFA.

Methods	*DSC*	*IoU*	*Pre*	*Sen*
Baseline	0.8154	0.6879	0.8749	0.7630
Baseline+nonlocal	0.8271	0.7052	0.8103	0.8447
Baseline+CSGM (level1, 4)	0.8457	0.7326	0.8932	0.8030
Baseline+CSGM (level2, 4)	0.8409	0.7255	0.8619	0.8209
Baseline+CSGM (level3, 4)	0.8358	0.7052	0.8103	0.8447
Baseline+CSGM (level1, 2, 4)	0.8278	0.7062	0.8955	0.7696
baseline+CSGM (level1, 3, 4)	0.8293	0.7083	0.8740	0.7888
Baseline+1LFA (level1–4)	0.8248	0.7018	0.8988	0.7620
Baseline+2LFA (level2–4)	0.8313	0.7112	0.8665	0.7988
baseline+3LFA (level3–4)	0.8353	0.7172	0.8898	0.7871
Ours(baseline+CSGM(level1,4)+2LFA(2-4))	**0.8498**	0.7389	0.8878	0.8149

Ablation Study. We investigate the effectiveness of the proposed CSGM and LFA. The results are shown in Table 2. The i-th level in the table means that CSGMs are added behind the i-th decoder blocks or the i-th decoder blocks are replaced with LFAs. We can conclude that both CSGM and LFA have improved the performance of the model to varying extent. Overall, placing two CSGM behind different levels of the model is better than placing three. It has achieved a 3.12%–3.72% improvement of the Dice score, clearly showing that the CSGM is necessary for increasing performance. Another conclusion that can be drawn is that the longer the model focus on the target region, the better the performance will be. The same ablation experiment was also conducted for LFA, different numbers of LFA added in different positions will all make improvements to our network (1.15%–2.44% of dice score). All the experiment results show the effectiveness of CSGM and LFA.

4 Conclusion

This study proposes a new architecture based on UNet for automatic CP segmentation. The network is formulated in a coarse-to-fine fashion, embedded with the Coarse Segmentation Guided Module (CSGM) and Local Feature Aggregation (LFA). CSGM can efficiently guide the model to only focus on the region of interest (the tumor and its surrounding tissues), and LFA can encode the fine-grained information and further mine the fine-grained cues. To validate the effectiveness of our model, a realistic CP dataset was constructed and the experiments suggest that this architecture can boost the performance both on boundary and contour without a dramatic increase in parameters.

Acknowledgements. This work was supported by the National Natural Science Fund for Distinguished Young Scholar under Grants No. 62025601.

References

1. Müller, H.L.: Childhood craniopharyngioma. Pituitary **16**(1), 56–67 (2013)
2. Stamm, A.C., Vellutini, E., Balsalobre, L.: Craniopharyngioma. Otolaryngol. Clin. North Am. **44**(4), 937–952 (2011)
3. Müller, H.L., Merchant, T.E., Warmuth-Metz, M., Martinez-Barbera, J.P., Puget, S.: Craniopharyngioma. Nature Rev. Disease Primers **5**(1), 1–19 (2019)
4. Inenaga, C., Kakita, A., Iwasaki, Y., Yamatani, K., Takahashi, H.: Autopsy findings of a craniopharyngioma with a natural course over 60 years. Surg. Neurol. **61**(6), 536–540 (2004)
5. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and ¡ 0.5 MB model size. arXiv preprint arXiv:1602.07360 (2016)
6. Fan, D., et al.: Inf-Net: automatic COVID-19 lung infection segmentation from CT images. IEEE Trans. Med. Imaging **39**(8), 2626–2637 (2020)
7. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-SCNN: gated shape CNNs for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5229–5238 (2019)
8. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
9. Zhang, Z., Fu, H., Dai, H., Shen, J., Pang, Y., Shao, L.: ET-Net: a generic Edge-aTtention guidance network for medical image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 442–450. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_49
10. Feng, S., et al.: CPFNet: context pyramid fusion network for medical image segmentation. IEEE Trans. Med. Imaging **39**(10), 3008–3018 (2020)
11. Kaluva, K.C., Khened, M., Kori, A., Krishnamurthi, G.: 2D-densely connected convolution neural networks for automatic liver and tumor segmentation. arXiv preprint arXiv:1802.02182 (2018)
12. Feng, X., Wang, C., Cheng, S., Guo, L.: Automatic liver and tumor segmentation of CT based on cascaded U-Net. In: Jia, Y., Du, J., Zhang, W. (eds.) Proceedings of 2018 Chinese Intelligent Systems Conference. LNEE, vol. 529, pp. 155–164. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-2291-4_16
13. Albishri, A.A., Shah, S.J.H., Lee, Y.: CU-Net: cascaded u-net model for automated liver and lesion segmentation and summarization. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1416–1423. IEEE (2019)
14. Yan, Y., et al.: Cascaded multi-scale convolutional encoder-decoders for breast mass segmentation in high-resolution mammograms. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6738–6741. IEEE (2019)
15. Ma, Q., Zu, C., Wu, X., Zhou, J., Wang, Y.: Coarse-to-fine segmentation of organs at risk in nasopharyngeal carcinoma radiotherapy. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 358–368. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_34
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)

18. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial Intelligence and Statistics, pp. 562–570. PMLR (2015)
19. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: Wang, Q., Shi, Y., Suk, H.-I., Suzuki, K. (eds.) MLMI 2017. LNCS, vol. 10541, pp. 379–387. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_44
20. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1
22. Jha, D., et al.: ResUNet++: an advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM), pp. 225–2255. IEEE (2019)
23. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
24. Ypsilantis, P.P., Montana, G.: Learning what to look in chest x-rays with a recurrent visual attention model. arXiv preprint arXiv:1701.06452 (2017)
25. Gu, R., et al.: CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging **40**(2), 699–711 (2020)
26. Fan, D.-P., et al.: PraNet: parallel reverse attention network for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 263–273. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_26
27. Song, J., et al.: Global and local feature reconstruction for medical image segmentation. IEEE Trans. Med. Imaging **41**(9), 2273–2284 (2022)

Patch-Level Instance-Group Discrimination with Pretext-Invariant Learning for Colitis Scoring

Ziang Xu[1,3], Sharib Ali[1,2(✉)], Soumya Gupta[1,3], Simon Leedham[4,5], James E. East[4,5], and Jens Rittscher[1,3,4]

[1] Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford, UK
s.s.ali@leeds.ac.uk
[2] School of Computing, University of Leeds, Leeds, UK
[3] Big Data Institute, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Oxford, UK
[4] NIHR Oxford Biomedical Research Centre, Oxford, UK
[5] Translational Gastroenterology Unit, Experimental Medicine Division, John Radcliffe Hospital, University of Oxford, Oxford, UK

Abstract. Inflammatory bowel disease (IBD), in particular ulcerative colitis (UC), is graded by endoscopists and this assessment is the basis for risk stratification and therapy monitoring. Presently, endoscopic characterisation is largely operator dependant leading to sometimes undesirable clinical outcomes for patients with IBD. We focus on the Mayo Endoscopic Scoring (MES) system which is widely used but requires the reliable identification of subltle changes in mucosal inflammation. Most existing deep learning classification methods cannot detect these fine-grained changes which make UC grading such a challenging task. In this work, we introduce a novel patch-level instance-group discrimination with pretext-invariant representation learning (PLD-PIRL) for self-supervised learning (SSL). Our experiments demonstrate both improved accuracy and robustness compared to the baseline supervised network and several state-of-the-art SSL methods. Compared to the baseline (ResNet50) supervised classification our proposed PLD-PIRL obtained an improvement of 4.75% on hold-out test data and 6.64% on unseen center test data for top-1 accuracy.

Keywords: Colonoscopy · Inflammation · Self-supervised learning · Classification · Group discrimination · Colitis

1 Introduction

Ulcerative colitis (UC) is a chronic intestinal inflammatory disease in which lesions such as inflammation and ulcers are mainly located in the colon and rectum. UC is more common in early adulthood, the disease lasts for a long time,

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_11.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 101–110, 2022.
https://doi.org/10.1007/978-3-031-21014-3_11

and the possibility of further cancerous transformation is high [15]. It is therefore important to diagnose ulcerative colitis early. Colonoscopy is a gold standard clinical procedure widely used for early screening of disease. Among the various colonoscopic evaluation methods proposed, the Mayo Endoscopic Score (MES) is considered to be the most widely used evaluation indicators to measure the UC activity [6,16]. MES divides UC into three categories, namely mild (MES-1), moderate (MES-2) and severe (MES-3). MES-2 and MES-3 indicate that an immediate follow up is required. However, the grading of UC in colonoscopy is dependent on the level of experience. Differences in assessment amongst endoscopists have been observed that can affect patient management. Automated systems based on artificial intelligence can help identify subtle abnormalities that represent UC, improve diagnostic quality and minimise subjectivity.

Deep learning models based on Convolutional Neural Networks (CNN) [9] have already been used to build UC MES scoring systems [2,11]. But rather than formulating the problem as a 3-way classification task that separates the three MES categories (mild, moderate, severe), existing methods resort to learning binary classifiers to deal with the high degree of intra-class similarity. Consequently, a number of different models needs to be trained.

We propose to amplify the classification accuracy of a CNN network for a 3-way classification using an invariant pretext representation learning technique in a self-supervised setting that exploits patch-based image transformations and additionally use these patches for instance-based group discrimination by grouping same class together using k-means clustering, referred to as "PLD-PIRL". The idea is to increase the intera-class separation and minimise the intra-class separation. The proposed technique uses a CNN model together with unsupervised k-means clustering for achieving this objective. The subtlety in the mucosal appearances are learnt by transforming images into a jigsaw puzzle and computing contrastive losses between feature embedding (*aka* representations). In addition, we also explore the introduction of an attention mechanism in our classification network to further boost classification accuracy. We would like to emphasise that the UC scoring is a complex classification task as image samples are very similar and often confusing to experts (especially between grades 1 and 2). Thus, developing an automated system for this task has tremendous benefit in clinical support system.

The related work on UC scoring based on deep learning in presented in Sect. 2. In Sect. 3, we provide details of the proposed method. Section 4 consists of implementation details, dataset preparation and results. Finally, a conclusion is presented in Sect. 5.

2 Related Work

2.1 CNN-Based UC Grading

Most research work on UC grading is based on MES scores. Mokter et al. [11] propose a method to classify UC severity in colonoscopy videos by detecting vascular (vein) patterns using three CNN networks and use a training dataset comprising of over 67k frames. The first CNN is used to discriminate between

a high and low density of blood vessels. Subsequently they use two CNNs separately for the subsequent UC classification each in binary two class configuration. Such a stacked framework can minimise false positives but does not enhance the model's ability to understand variability of different MES scores. Similarly, Stidham et al. [14] use the Inception V3 model to train and evaluate MES scores in still endoscopic frames. They used 16k UC images and obtained an accuracy of 67.6%, 64.3% and 67.9% for the three MES classes. UCS-CNN [1] includes several prepossessing steps such as interpretation of vascular patterns, patch extraction techniques and CNN-based classifier for classification. A total of 92000 frames were used for training obtaining accuracy of 53.9% 62.4% and 78.9% for mild, moderate and severe classes. Similarly, Ozawa et al. [13] use a CNN for binary classification only on still frames comprising of 26k training images, which first between normal (MES 0) and mucosal healing state (MES 1) while next between moderate (MES 2) and severe (MES 3). Gutierrez et al. [2] used CNN model to predict a binary version of the MES scoring of UC.

One common limitation of existing CNN-based UC scoring literature is the use of existing multiple CNN models in an ensemble configuration, simplifying MES to a binary problem and use of very large in-house datasets for training. In contrast, we aim to develop a single CNN model-based approach for a 3-way MES scoring that is clinically relevant. In addition, we use a publicly available UC dataset [3] to guarantee reproducibility of our approach. Furthermore, this dataset consists of only 851 image samples and therefore poses a small data problem.

2.2 Self-supervised Approach for Classification

Self-supervised learning (SSL) uses pretext tasks to mine self-supervised information from large-scale unsupervised data, thereby learning valuable image representations for downstream tasks. By doing so, the limitation of network performance on predefined annotations are greatly reduced. In SSL, the pretext task typically applies a transformation to the input image and predicts properties of the transformation from the transformed image. Chen *et al.* [4] proposed the SimCLR model, which performs data enhancement on the input image to simulate the input from different perspectives of the image. Contrastive loss is then used to maximize the similarity of the same object under different data augmentations and minimised the similarity between similar objects. Later, the MoCo model proposed by He et al. [7] also used contrastive loss to compare the similarity between a query and the keys of a queue to learn feature representation. The authors used a dynamic memory, rather than static memory bank, to store feature vectors used in training. In contrast to these methods that encourages the construction of covariant image representations to the transformations, pretext-invariant representation learning (PIRL) [10] pushes the representations to be invariant under image transformations. PIRL computes high similarity to the image representations that are similar to the representation of the transformed versions and low similarity to representations for the different images. Jigsaw puzzle [12] was used as pretext task for PIRL representation learning.

Inspired by PIRL, we propose a novel approach that exploits the invariant representation learning together with patch-level instance-group discrimination.

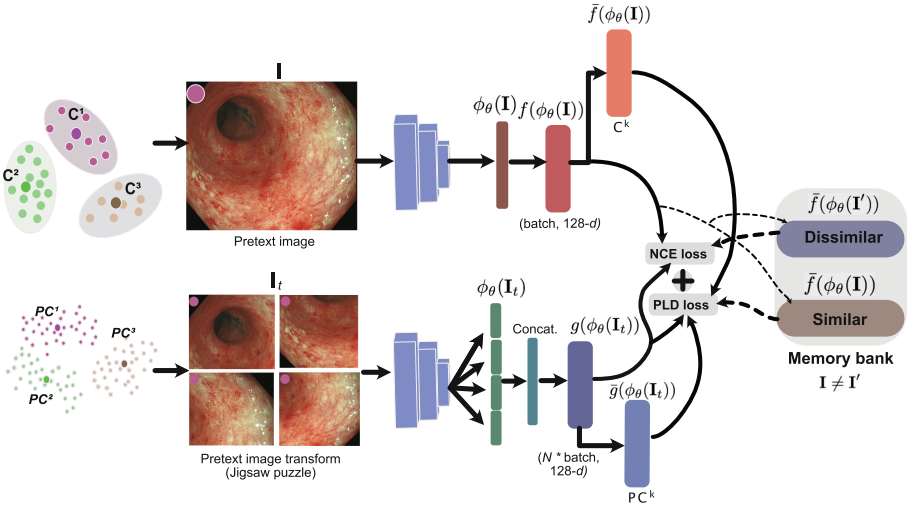


Fig. 1. Pretext invariant patch-level instance group discrimination for ulcerative colitis (UC) scoring. Two identical classification networks are used to compute image-level and patch-level embedding. Three Mayo Endoscopic Scoring (MES) for UC from 1 up to 3 (mild: 1, moderate: 2 and severe: 3) are presented as three separate clusters for both images and patches. The memory bank contains the moving average of representations for all images in the dataset. Here, $\mathbf{I}$ represent an image sample while $\mathbf{I}_t$ is a transformed puzzle of that image and $\mathbf{I}'$ represent negative sample.

Here, the idea is to increase the inter-class separation and minimise the intra-class separation. An unsupervised k-means clustering is used to define feature clusters for k-class categories. We demonstrate the effectiveness of this approach on ulcerative colitis (UC) dataset. UC scoring remains a very challenging classification task, while being very important task for clinical decision making and minimising current subjectivity.

3 Method

We propose to increase inter-class separation and minimise the intra-class distance by jointly minimising two loss functions that are based on contrastive loss. In contrast to widely used image similarity comparisons we use patch-level and image-level configurations. Additionally, we propose a novel instance group level discriminative loss. A memory bank is used to store moving average embedding of negative samples for efficient memory management. The block diagram of our proposed MES-scoring for ulcerative colitis classification framework is shown in Fig. 1.

3.1 Pretext Invariant Representations

Let the ulcerative colitis dataset consists of N image samples denoted as $\mathcal{D}_{uc} = \{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N\}$ for which a transformation $\mathcal{T}$ is applied to create and reshuffle m number of image patches for each image in $\mathcal{D}_{uc}$, $\mathcal{P}_{uc} =$

$\{\mathbf{I}_{1t}^1, ..., \mathbf{I}_{1t}^m, ..., \mathbf{I}_{Nt}^1, ..., \mathbf{I}_{Nt}^m\}$ with $\mathcal{T} \in t$. We train a convolutional neural network with free parameters θ that embody representation $\phi_\theta(\mathbf{I})$ for a given sample $\mathbf{I}$ and $\phi_\theta(\mathbf{I}_t)$ for patch $\mathbf{I}_t$. For image patches, representations of each patch constituting the image $\mathbf{I}$ is concatenated. A unique projection heads, $f(.)$ and $g(.)$, are applied to re-scale the representations to a 128-dimensional feature vector in each case (see Fig. 1). A memory bank is used to store positive and negative sample embedding of a mini-batch B (in our case, $B = 32$). Negative refer to embedding for $I' \neq I$ that is required to compute our contrastive loss function $\mathcal{L}(.,.)$, measuring the similarity between two representations. The list of negative samples, say $\mathcal{D}_n$, grows with the training epochs and are stored in a memory bank $\mathcal{M}$. To compute a noise contrastive estimator (NCE), each positive samples has $|\mathcal{D}_n|$ negative samples and minimizes the loss:

$$\mathcal{L}_{NCE}(\mathbf{I}, \mathbf{I}_t) = -\log[h(f(\phi_\theta(\mathbf{I})), g(\phi_\theta(\mathbf{I}_t)))] - \sum_{\mathbf{I}' \in \mathcal{D}_n} \log[1 - h(f(\phi_\theta(\mathbf{I}')), g(\phi_\theta(\mathbf{I}_t)))]. \tag{1}$$

For our experiments we have used both ResNet50 [8] and a combination of convolutional block attention (CBAM, [17]) with ResNet50 model (ResNet50^{+cbam}) for computing the representation $f(.)$ and $g(.)$. In Eq. (1), $h(.,.)$ is the *cosine* similarity between the representations with a temperature parameter τ, and for $h(f(.), g(.))$:

$$h(f, g) = \frac{\exp < f, g > /\tau}{\exp < f, g > /\tau + |\mathcal{D}_n|/N}. \tag{2}$$

The presented loss encourages the representation of image $\mathbf{I}$ to be similar to its corresponding transformed patches $\mathbf{I}_t$ while increasing the distance between the dissimilar image samples $\mathbf{I}'$. This enables network to learn invariant representations.

3.2 Patch-Level Instance Group Discrimination Loss

Let $\bar{f}(.)$ and $\bar{g}(.)$ be the mean embedding for classes k with cluster centers C^k and PC^k respectively for the image $\mathbf{I}$ and patch samples $\mathbf{I}_t$. k-means clustering is used to group the embedding into k $(= 3)$ class instances. The idea is to then compute the similarity of each patch embedding $g(.)$ with the mean image embedding $\bar{f}(.)$ for all k classes and vice-versa using Eq. (2). A cross-entropy (CE) loss is then computed that represent our proposed $L_{PLD}(.,.)$ loss function given as:

$$\begin{aligned} \mathcal{L}_{PLD}(\mathbf{I}, \mathbf{I}_t) = &-0.5 \sum_{\forall k} C^k \log(h(\bar{f}(\phi_\theta(\mathbf{I})), g(\phi_\theta(I_t))) \\ &-0.5 \sum_{\forall k} PC^k \log(h(\bar{g}(\phi_\theta(\mathbf{I}_t)), f(\phi_\theta(\mathbf{I}))). \end{aligned} \tag{3}$$

The proposed patch-level group discrimination loss $\mathcal{L}_{PLD}(\mathbf{I}, \mathbf{I}_t)$ takes k class instances into account. As a result not only the similarity between the group (mean) embedding and a single sample embedding for the same class is maximised but also it guarantees inter-class separation. The final loss function with

empirically set $\lambda = 0.5$ is minimised in our proposed PLD-PIRL network and is given by:

$$\mathcal{L}_{final}(\mathbf{I}, \mathbf{I}_t) = \mathcal{L}_{NCE}(\mathbf{I}, \mathbf{I}_t) + \lambda \mathcal{L}_{PLD}(\mathbf{I}, \mathbf{I}_t). \quad (4)$$

4 Experiments and Results

4.1 Implementation Details

For training of pretext tasks in self-supervised learning, we use a learning rate (LR) of $1e^{-3}$ and an SGD optimizer. 3000 epochs with a batch size of 32 were used to train pretext tasks presented in all experiments. For PLD loss, we set $k = 3$ for a number of clusters and test the effect of different temperatures τ and λ on the model performance. For the next downstream classification task, we finetune the model with the LR of $1e^{-4}$, the SGD optimizer, batch size of 32, and the LR decay of 0.9 times per 30 epochs. Our experimental results showed that most of the models converged around 150 epochs. For the baseline, supervised model training converged to higher epochs (nearly 200). The stopping criteria were based on minimal loss improvement of 0.000001 over 20 consecutive epochs. The proposed method is implemented on a server deployed with an NVIDIA Quadro RTX 6000 graphics card using the PyTorch framework. All input images were resized to 224×224 pixels.

4.2 Datasets and Evaluation

We have used both publicly available and in-house dataset. HyperKvasir [3] public dataset was used for model training, validation and as hold-out test samples (referred as Test-I). The available dataset includes MES scores (1,2 and 3) and three additional scoring levels categorising into scores 0-1, 1-2 and 2-3, totaling to 6 UC categories and 851 images. After re-examination by expert colonoscopist, the final data was divided into three different grades: mild, moderate and severe. In the experiment, 80% of the data is used for training, 10% for validation and 10% for testing. Furthermore, to evaluate the efficacy of the proposed PLD-PIRL method on unseen center data (Test-II), we used one in-house dataset. This dataset contains 151 images from 70 patient videos. We manually selected frames containing UC from the videos, which were then labeled as mild, moderate, and severe by an expert colonoscopist.

We have used standard top-k accuracy (percentage of samples predicted correctly), F1-score ($= \frac{tp}{tp+fp}$, tp: true positive, fp: false positive), specificity ($= \frac{tp}{tp+fn}$) and sensitivity ($= \frac{tn}{tn+fp}$) and for our 3-way classification task of MES-scoring for UC.

4.3 Comparison with SOTA Methods

Result of baseline fully supervised classification and self-supervised learning model (SSL) for UC classification on two test datasets (Test-I and Test-II)

Table 1. Experimental results on hold-out test set (Test-I) and unseen center test set (Test-II) for a three way classification of MES scores (1, 2 and 3). Two best results are in bold. Both classical supervised (baseline) and self-supervised methods are compared with our proposed PLD-PIRL classification method.

Method	Model	Top 1	Top 2	F1	Sen.	Spec.
Test-I						
Baseline	ResNet50	64.29%	91.63%	62.77%	60.56%	81.02%
	ResNet50^{+cbam}	65.47%	93.59%	64.38%	63.58%	81.95%
SimCLR [4]	ResNet50	61.91%	94.93%	59.87%	58.49%	79.60%
SimCLR [4]	ResNet50^{+cbam}	63.09%	92.27%	60.90%	59.31%	79.84%
SimCLR + DCL [5]	ResNet50	64.28%	**94.98%**	62.34%	62.78%	80.35%
SimCLR + DCL [5]	ResNet50^{+cbam}	64.79%	94.06%	62.01%	62.39%	79.56%
MOCO + CLD [7]	ResNet50	66.96%	93.12%	65.79%	66.38%	84.26%
MOCO + CLD [7]	ResNet50^{+cbam}	67.32%	92.52%	66.38%	65.91%	**84.53%**
PIRL [10]	ResNet50	65.93%	92.89%	64.87%	64.26%	81.03%
PIRL [10]	ResNet50^{+cbam}	66.67%	93.21%	65.91%	66.47%	82.59%
PLD-PIRL (ours)	ResNet50	**67.85%**	93.98%	**67.48%**	**66.93%**	83.41%
PLD-PIRL(ours)	ResNet50^{+cbam}	**69.04%**	**96.31%**	**68.98%**	**67.35%**	**84.71%**
Test-II						
Baseline	ResNet50	57.61%	88.36%	57.03%	56.69%	71.10%
	ResNet50^{+cbam}	60.92%	90.49%	59.38%	58.81%	73.79%
SimCLR [4]	ResNet50	56.95%	85.88%	55.91%	54.69%	71.26%
SimCLR [4]	ResNet50^{+cbam}	57.31%	85.21%	56.29%	56.50%	71.98%
SimCLR + DCL [5]	ResNet50	58.94%	87.92%	57.36%	57.29%	73.29%
SimCLR + DCL [5]	ResNet50^{+cbam}	59.60%	90.34%	58.42%	59.58%	74.19%
MOCO + CLD [7]	ResNet50	60.61%	90.71%	60.52%	59.88%	75.69%
MOCO + CLD [7]	ResNet50^{+cbam}	60.93%	92.12%	60.61%	59.29%	77.33%
PIRL [10]	ResNet50	61.59%	92.61%	60.55%	60.53%	75.98%
PIRL [10]	ResNet50^{+cbam}	62.25%	**93.92%**	61.96%	60.92%	78.03%
PLD-PIRL (ours)	ResNet50	**62.90%**	92.93%	**62.81%**	**61.79%**	**80.23%**
PLD-PIRL(ours)	ResNet50^{+cbam}	**64.24%**	**95.32%**	**64.38%**	**62.99%**	**80.09%**

Sen. - sensitivity; Spec. - specificity

are presented in Table 1. ResNet50 and ResNet50^{+cbam} are established as the baseline model for supervised learning and the same are also used for other state-of-the-art (SOTA) SSL comparisons. In Table 1 for Test-I dataset, it can be observed that the proposed PLD-PIRL approach using ResNet50^{+cbam} model achieves the best results with 69.04%, 68.98%, 84.71% and 67.35%, respectively, for top 1 accuracy, F1 score, specificity and sensitivity. Compared to the supervised learning based baseline models i.e., ResNet50 and ResNet50^{+cbam}, the top 1 accuracy is improved by 4.75% and 3.57%, respectively, for these models using our proposed PLD-PIRL. We also compared the proposed PLD-PIRL approach with other SOTA self-supervised learning methods including popular

SimCLR [4], SimCLR+DCL [18], MOCO+CLD [7] and PIRL [10] methods. Our proposed network (ResNet50) clearly outperforms all these methods with at least nearly 1.1% (MOCO+CLD) up to 6% (SimCLR) on top-1 accuracy.

Similarly, for out-of-sample unseen center Test-II dataset (see Table 1), the proposed model outperforms the baseline fully supervised models by a large margin accounting to nearly 5% for ResNet50 and 4% for ResNet50^{+cbam}. A similar trend is observed for all SOTA SSL methods ranging from 3.63% for MOCO+CLD upto 5.3% for SimCLR with ResNet50. A clear boost of 1.99% can be seen for the best PLD-PIRL (ResNet50^{+cbam}) model compared to the PIRL (ResNet50^{+cbam}). Our experiments on all the existing approaches with ResNet and CBAM (ResNet^{+cbam}) backbone showed nearly 1% improvement over ResNet50 on test-I with other SOTA (e.g., top 1 accuracies for SimCLR: 63.09%, SimCLR+DCL: 64.79% and for MOCO+CLD: 67.32%) and around 0.5% on test-II (SimCLR: 57.31%, SimCLR+DCL: 59.60% and for MOCO+CLD: 60.93%). Figure 2 (left) representing t-SNE plots demonstrate an improved separation of sample points in both training and test sets compared to fully supervised baseline approach. Confusion matrix for both Test-I and Test-II are provided in supplementary Fig. 1 that shows that proposed PLD-PIRL were able to classify more samples compared to the baseline method. Similarly, it can be observed from supplementary Fig. 2 that the wrongly classified ones are only between adjacent classes which at times categories as both classes by the clinical endoscopists.

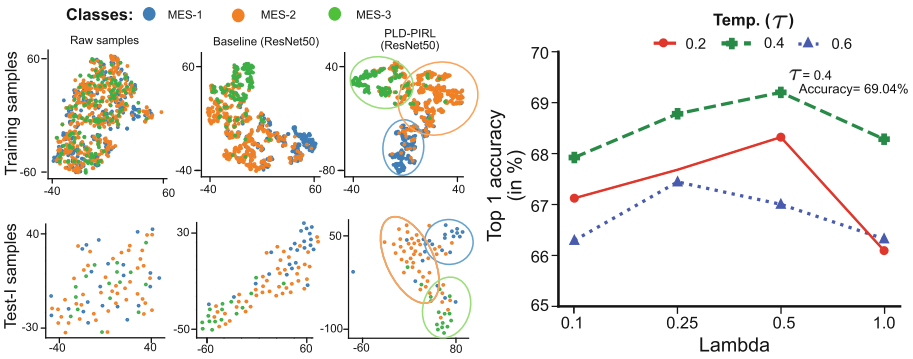


Fig. 2. (Left) Classified clusters for three MES classes obtained from fully supervised baseline and proposed PLD-PIRL on both training (top) and test-I (down) samples. Raw sample distributions are also shown. A t-distributed stochastic neighbor embedding is used for the point plots of image samples embedding. (Right) Experiments for finding best values for hyper-parameters temperature τ and λ weights in the loss function.

4.4 Ablation Study

Our experiments indicate that the settings of λ and temperature τ parameters in the proposed PLD-PIRL approach will affect the model performance. Therefore,

we conducted an ablation study experiment to further study the performance of PLD-PIRL under different parameter settings. We set $\tau = \{0.2, 0.4, 0.6\}$ and $\lambda = \{0.1, 0.25, 0.5, 1.0\}$. As can it can be observed from the plot in Fig. 2 (right) that for $\tau = 0.4$ PLD-PIRL maintained high accuracy at different λ values, and is better than other parameter settings. The best value is obtained at $\lambda = 0.5$ with the top1 accuracy of 69.04%.

5 Conclusion

Our novel self-supervised learning method using pretext-invariant representation learning with patch-level instance-group discrimination (PLD-PIRL) applied to the UC classification task overcomes the limitations of previous approaches that rely on binary classification tasks. We have validated our method on a public dataset and an unseen dataset. Our experiments show that compared with other SOTA classification methods that include fully supervised baseline models, our proposed method obtained large improvements in all metrics. The test results on the unseen dataset provides an evidence that our proposed PLD-PIRL method can learn to capture the subtle appearance of mucosal changes in colonic inflammation and the learnt feature representations together with instance-group discrimination allows improved accuracy and robustness for clinically use Mayo Endoscopic Scoring of UC.

References

1. Alammari, A., Islam, A.R., Oh, J., Tavanapong, W., Wong, J., De Groen, P.C.: Classification of ulcerative colitis severity in colonoscopy videos using CNN. In: Proceedings of the 9th International Conference on Information Management and Engineering, pp. 139–144 (2017)
2. Becker, B.G., et al.: Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data. Therap. Adv. Gastrointest. Endosc. **14** (2021)
3. Borgli, H., et al.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data **7**(1), 1–14 (2020)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
5. Chuang, C.Y., Robinson, J., Lin, Y.C., Torralba, A., Jegelka, S.: Debiased contrastive learning. Adv. Neural. Inf. Process. Syst. **33**, 8765–8775 (2020)
6. D'haens, G., et al.: A review of activity indices and efficacy end points for clinical trials of medical therapy in adults with ulcerative colitis. Gastroenterology **132**(2), 763–786 (2007)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

9. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
10. Misra, I., Maaten, L.V.D.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6707–6717 (2020)
11. Mokter, M.F., Oh, J.H., Tavanapong, W., Wong, J., de Groen, P.C.: Classification of ulcerative colitis severity in colonoscopy videos using vascular pattern detection. In: Liu, M., Yan, P., Lian, C., Cao, X. (eds.) MLMI 2020. LNCS, vol. 12436, pp. 552–562. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59861-7_56
12. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
13. Ozawa, T., et al.: Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. Gastrointest. Endosc. **89**(2), 416–421 (2019)
14. Stidham, R.W., et al.: Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. JAMA Netw. Open **2**(5), e193963–e193963 (2019)
15. Torres, J., et al.: Results of the seventh scientific workshop of ECCO: precision medicine in IBD-prediction and prevention of inflammatory bowel disease. J. Crohn's Colitis **15**(9), 1443–1454 (2021)
16. Vashist, N.M., et al.: Endoscopic scoring indices for evaluation of disease activity in ulcerative colitis. Cochrane Database Syst. Rev. (1) (2018)
17. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
18. Yeh, C.H., Hong, C.Y., Hsu, Y.C., Liu, T.L., Chen, Y., LeCun, Y.: Decoupled contrastive learning. arXiv preprint arXiv:2110.06848 (2021)



AutoMO-Mixer: An Automated Multi-objective Mixer Model for Balanced, Safe and Robust Prediction in Medicine

Xi Chen[1], Jiahuan Lv[1], Dehua Feng[1], Xuanqin Mou[1], Ling Bai[1], Shu Zhang[1], and Zhiguo Zhou[2(✉)]

[1] School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, China
xi_chen@mail.xjtu.edu.cn

[2] Department of Biostatistics and Data Science, University of Kansas Medical Center, Kansas City, KS, USA
zgzhou2013@gmail.com

Abstract. Accurately identifying patient's status through medical images plays an important role in diagnosis and treatment. Artificial intelligence (AI), especially the deep learning, has achieved great success in many fields. However, more reliable AI model is needed in image guided diagnosis and therapy. To achieve this goal, developing a balanced, safe and robust model with a unified framework is desirable. In this study, a new unified model termed as automated multi-objective Mixer (AutoMO-Mixer) model was developed, which utilized a recent developed multiple layer perceptron Mixer (MLP-Mixer) as base. To build a balanced model, sensitivity and specificity were considered as the objective functions simultaneously in training stage. Meanwhile, a new evidential reasoning based on entropy was developed to achieve a safe and robust model in testing stage. The experiment on an optical coherence tomography dataset demonstrated that AutoMO-Mixer can obtain safer, more balanced, and robust results compared with MLP-Mixer and other available models.

Keywords: Image guided diagnosis and therapy · Reliable artificial intelligence · Balance · Safe · Robustness

1 Introduction

With the development of modern medicine, medical image has become an essential tool to carry out personalized and accurate diagnosis. Due to the strong ability to analyze image, deep learning has been widely used in medical image analysis and has achieved great success [1,2] in the past years. However, many current available models can also lead to unreliable predictions. For example, the car's perception system misclassified the white part of the trailer into the sky, resulting in a fatal accident [3]. As such, different from other application

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 111–120, 2022.
https://doi.org/10.1007/978-3-031-21014-3_12

fields such as face recognition, nature image classification, model reliability is more important in medicine as it is related to human life and health. On the other hand, we not only need to obtain the accurate prediction results, but also need to know whether the outcome is reliable or not. To realize this abstract goal by considering the clinical needs, we believe that building a unified model to achieve balance, safe and robust is desirable.

Currently, most prediction models use a single objective (e.g., accuracy, AUC) [4,5] function in the model training. However, the imbalanced sensitivity and specificity may result in higher rate of missed diagnosis [6]. Therefore, a multi-objective model which considers sensitivity and specificity simultaneously is needed. So far, there have been some studies on multi-objective optimization [7,8].

Furthermore, since most models are data-driven based strategy, it is hard to evaluate whether the prediction outcome for an unseen sample is reliable or not. A possible solution is evaluating the model output by introducing a "third party" to independently estimate the model reliability or uncertainty. There have been several studies on uncertainty estimation for deep learning. [9] proposed a framework based on test-time data augmentation to quantify the diagnostic uncertainty in deep neural networks. [10] used the prediction of the augmented images to obtain entropy to estimate uncertainty.

Meanwhile, it is found that the model built based on the dataset collected from one institution always obtain bad performance when the testing dataset is from another institution [11–13], demonstrating the poor robustness. On the other hand, a reliable model should always work well across the multiple institutions. Several studies have investigated this issue. Adversarial attack is one of the most serious factors that cause models not to be robust [14]. Some attackers perturbated test reports to obtain medical compensation [15]. Adversarial examples lead to wrong decisions that can cause dangerous effects on the patient's life [16]. [17,18] evaluated the robustness of the model with adversarial attacks.

In summary, there have been several studies on building balanced, safe and robust model independently, but there is no unified framework that can achieve three goals simultaneously. As such, a new automated multi-objective Mixer (AutoMO-Mixer) model based on multiple layer perceptron Mixer (MLP-Mixer) is developed in this study to build a more reliable model. In AutoMO-Mixer, both sensitivity and specificity were considered as the objective functions simultaneously and a Pareto-optimal model set can be obtained through the multi-objective optimization [20] in training stage. In testing stage, the Pareto-optimal models with balanced sensitivity and specificity were chosen so as to improve model balance. To obtain safer and more robust model, evidential reasoning based on entropy (ERE) approach was developed to fuse the outputs of Pareto-optimal models to obtain the final outcome. The experimental studies on optical coherence tomography (OCT) dataset demonstrated that AutoMO-Mixer can outperform MLP-Mixer and other deep learning models, and more balanced, robust and safer results can be achieved as well.

2 Method

2.1 Overview

The framework of AutoMO-Mixer is shown in Fig. 1, which consists of training and testing stages. To build a balanced model, both sensitivity and specificity are considered as objective functions simultaneously in training stage, and a Pareto-optimal model set is generated then. To build a safer and more robust model, ERE strategy is developed to fuse the probability outputs of multiple Pareto-optimal models in testing stage.

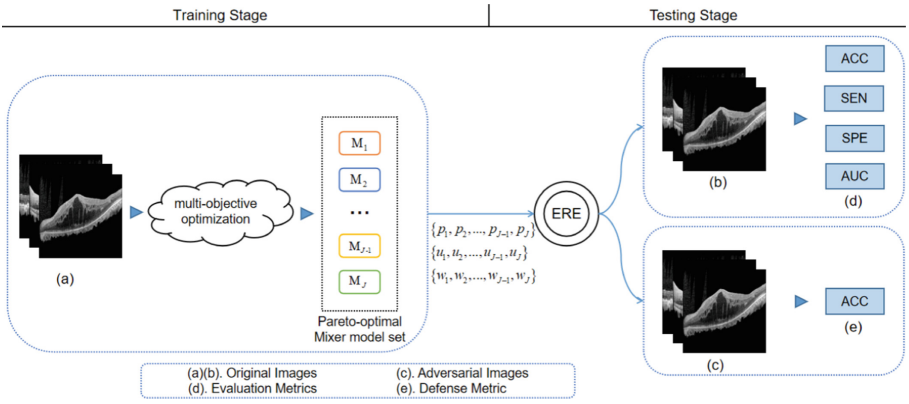


Fig. 1. The framework of AutoMO-Mixer model.

2.2 MLP-Mixer

Since the computational complexity is increased sharply when there are more parameters in multi-objective learning, it is better to have fewer parameters in model training. The recently proposed MLP-Mixer [19] model is a full MLP architecture. Compared with CNN, the convolutional layer is removed from MLP-Mixer, leading to decreasing the scale of the architecture parameters sharply. On the other hand, MLP-Mixer can achieve similar performance to CNN [19]. Therefore, it is a better choice in multi-objective learning.

2.3 Training Stage

In training stage, sensitivity denoted by f_{spe} and specificity denoted by f_{sen} are considered as objective functions simultaneously, they are:

$$f_{sen} = \frac{TP}{TP + FN} \tag{1}$$

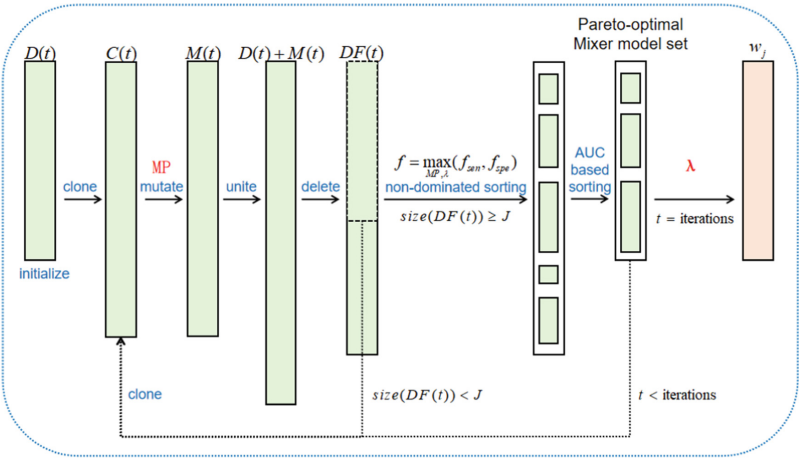


Fig. 2. The illustration of training stage.

$$f_{spe} = \frac{TN}{TN + FP} \tag{2}$$

where TP and TN represent the number of true positives and true negatives, FP and FN are the number of false positives and false negatives, respectively.

Assume M $= \{m_1, ..., m_q\}$ denotes the MLP-Mixer model, where q represents the number of model parameters. To obtain the balanced models, we aim to maximize f_{sen}, f_{spe} simultaneously, and an iterative multi-objective immune algorithm (IMIA) [20] is used. IMIA consists of six steps: initialization, cloning, mutation, deletion, update, and termination. First, the initial model set denoted by $D(t) = \{\mathrm{M}_1, ..., \mathrm{M}_N\}$ is generated , where $\mathrm{M}_i = \{m_{i1}, ..., m_{iq}\}, i = 1, 2, ..., N$. Then the models with higher f_{sen}, f_{spe} will be replicated using the proportional cloning method. In the third step, a probability of mutation is randomly generated for each model, and the model performs mutation when its probability is larger than the mutation probability (MP). After the mutation, the new models are generated. If some models have same sensitivity and specificity, only one model is remained. Then the model set size is kept through AUC based non-dominated sorting strategy. The training process will not stop until the maximum number of iterations is reached. Finally, the Pareto-optimal Mixer model set is generated, where the model set size is J. Since the two hyperparameters MP and λ may affect the model performance, Bayesian optimization [21] is used to optimize the hyperparameters. The illustration of the training phase is shown in Fig. 2.

2.4 Testing Stage

In testing stage, the probability outputs of Pareto-optimal models are fused through the evidential reasoning [22,23] based on entropy approach. The workflow is shown in Fig. 3.

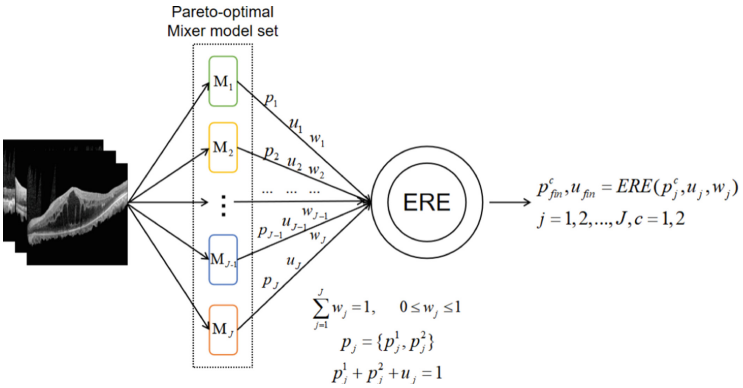


Fig. 3. The illustration of testing stage.

Weight Calculation. Since the performance of different Pareto-optimal models cannot be the same, the weight for each model should be estimated, which is denoted by w_j. As the balanced model between sensitivity and specificity is desired, the ratio between them is considered in the weight calculation, that is $\frac{f_{sen}}{f_{spe}}$ or $\frac{f_{spe}}{f_{sen}}$. When the ratio is less than 0.5 or greater than 1, the model is considered as extreme imbalance, setting w_j as 0. Meanwhile, AUC is a good measure for model reliability, it is also considered. The expression of w_j is as follows:

$$w_j = \begin{cases} \lambda \frac{f^j_{sen}}{f^j_{spe}} + (1-\lambda) AUC_j, & when\ 0.5 \le \frac{f^j_{sen}}{f^j_{spe}} \le 1 \\ \lambda \frac{f^j_{spe}}{f^j_{sen}} + (1-\lambda) AUC_j, & when\ 0.5 \le \frac{f^j_{spe}}{f^j_{sen}} \le 1, j = 1,2,...,J \\ 0 & Other\ situations \end{cases} \tag{3}$$

where λ indicates the importance of balance, and $1-\lambda$ indicates the importance of AUC. After calculating the w_j for each model, the weights are normalized.

Uncertainty Estimation. Test-time data augmentation (TTA) [9,10] is used to perform useful estimates of model uncertainty. The test image is fed into model $M_j, j = 1,2,...,J$ to generate the probability output $p^c_j, c = 1,2$, where $p^1_j + p^2_j = 1$. The original test image is enhanced T times to generate prediction $p^c_{j,t}, t = 1,2,...,T$. The mean class probability $\overline{p^c_j}$ and the uncertainty u_j are:

$$\overline{p^c_j} = \frac{1}{T} \sum_{t=1}^{T} p^c_{j,t}\ , c = 1,2 \tag{4}$$

$$u_j = -\sum_{c=1}^{2} \overline{p^c_j} \log(\overline{p^c_j}) \tag{5}$$

To satisfy the conditions of the ERE strategy, p^c_j and u_j are normalized so that $p^1_j + p^2_j + u_j = 1$.

ERE Strategy. Assume that the output probability for each model is denoted by $p_j = \{p_j^1, p_j^2\}, p_j^1 + p_j^2 \leq 1, j = 1, 2, ..., J$. If $p_j^1 + p_j^2 < 1$, it shows that the jth model has uncertainty u_j on its output. Then the final output probability $p_{fin}^c, c = 1, 2$ and uncertainty u_{fin} are obtained through the ERE fusion strategy. that is:

$$p_{fin}^c, u_{fin} = ERE(p_j^c, u_j, w_j), j = 1, 2, ..., J, c = 1, 2 \tag{6}$$

where ERE is:

$$p_{fin}^c = \frac{\mu \times [\prod_{j=1}^{J} (w_j p_j^c + 1 - w_j(p_j^1 + p_j^2)) - \prod_{j=1}^{J} (1 - w_j(p_j^1 + p_j^2))]}{1 - \mu \times [\prod_{j=1}^{J} (1 - w_j)]}, c = 1, 2 \tag{7}$$

$$u_{fin} = \frac{\mu \times [\prod_{j=1}^{J} (1 - w_j(p_j^1 + p_j^2)) - \prod_{j=1}^{J} (1 - w_j)]}{1 - \mu \times [\prod_{j=1}^{J} (1 - w_j)]} \tag{8}$$

The normalized factor μ is:

$$\mu = [\sum_{c=1}^{2} \prod_{j=1}^{J} (w_j p_j^c + 1 - w_j(p_j^1 + p_j^2)) - \prod_{j=1}^{J} (1 - w_j(p_j^1 + p_j^2))]^{-1} \tag{9}$$

2.5 Robustness Evaluation

In this study, fast gradient sign method (FGSM) [24] is used to disturb the original samples, which is a white box attack with full information of the models. Adversarial samples are generated by the following formula:

$$x^a = x + \delta \tag{10}$$

where x^a represents the adversarial sample, x represents the original sample. δ represents the perturbation. The degree of perturbation is controlled by ε. In our study, ACC is used to evaluate robustness [18].

3 Experiments

3.1 Experimental Setup

The dataset used in this study was collected from the Second Affiliated Hospital of Xi'an Jiaotong University (Xi'an, China), including 228 patients with Choroidal neovascularization (CNV) and cystoid macular edema (CME) between October 2017 and October 2019. First, OCT images of each patient were acquired via the Heidelberg Retina Tomograph-IV (Heidelberg Engineering, Heidelberg, Germany). These patients were then injected with anti-vascular endothelial

Table 1. The range of values for MLP-Mixer network structure parameters.

Parameters	Model parameter range
Number of layers	[2, 3, 4]
Hidden size C	256*[1, 1.2, 1.4, 1.6]
MLP dimension Ds	196*[2, 3, 4, 5]
MLP dimension Dc	256*[2, 4, 6, 8, 10]

Table 2. The evaluation results on OCT dataset.

Models	SEN	SPE	AUC	ACC	$\frac{\min(SEN,SPE)}{\max(SEN,SPE)}$
MLP-Mixer	0.611 ± 0.052	0.703 ± 0.077	0.709 ± 0.041	0.671 ± 0.038	0.869
ResNet-18	0.728 ± 0.075	0.706 ± 0.071	0.791 ± 0.046	0.714 ± 0.052	0.970
AutoMO-Mixer	0.778 ± 0.000	0.779 ± 0.000	0.844 ± 0.000	0.779 ± 0.000	**0.999**

growth factor (anti-VEGF) and the evaluations were made after 21 days. Among them, anti-VEGF was effective for 171 patients, and the remaining 57 patients had no sign of effectiveness. The study was approved by the Research Ethics Committee, and each patient provided written informed consent. In our study, we built a binary classifier to determine whether anti-VEGF would be effective for patients using OCT images. In the training stage, there were 135 effective cases and 44 ineffective cases. In the testing stage, there were 34 and 12, respectively, in these two classes.

Before being fed into the model, all the images were resized into 224×224. MP and λ were set to 0.5 and 0.8, respectively. The MLP-Mixer contains four parameters, these settings are shown in Table 1.

As AutoMO-Mixer was built based on MLP-Mixer and ResNet-18 is a classical deep learning model, they were used in comparative study. The four parameters in MLP-Mixer network were set to 5, 256, 392, 1024, respectively, and transfer learning was used on ResNet-18 as pre-training. Sensitivity (SEN), specificity (SPE), area Under Curve (AUC), and accuracy (ACC) were used for evaluation. All the experiments were performed five times, and mean and standard deviation were evaluated.

3.2 Results

The evaluation results on MLP-Mixer, ResNet-18 and AutoMO-Mixer are shown in Table 2. In this study, $\frac{\min(SEN,SPE)}{\max(SEN,SPE)}$ was used to assess the balance of the model. It can be seen that AutoMO-Mixer model is the most balanced. In addition, both the AUC and ACC of the AutoMO-Mixer are better than the other two models.

Safety Evaluation. In this study, the uncertainty estimation was used to measure model safety. If the performance of the model can improve as the number

Table 3. Model performance of the test cohorts stratified by the uncertainty.

Uncertainty	SEN	SPE	AUC	ACC
0.4245	0.778	0.779	0.844	0.779
0.4206	0.783	0.796	0.860	0.792
0.4165	0.818	0.829	0.823	0.827
0.4045	1.000	0.895	1.000	0.920

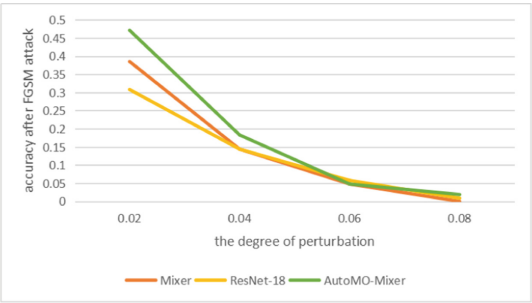


Fig. 4. Comparison of the robustness between the AutoMO-Mixer, ResNet-18 and AutoMO-Mixer models.

of test samples with high uncertainty decreases, it is indicated that the model is safe. The entire test samples were arranged from smallest to largest in order of uncertainty, with the maximum uncertainty being 0.4245, the upper quartile being 0.4206, the median being 0.4165, and the lower quartile being 0.4045. Samples with less uncertainty than them were grouped into four cohorts, and the evaluation results are shown in Table 3. It can be seen that the lower the cutoff uncertainty is, the better the model's performance is, indicating our model can assess whether the prediction is safe based on uncertainty.

Robustness. After the original samples were attacked by FGSM, indistinguishable adversarial samples were generated. We measured the accuracy of adversarial samples in each model in Fig. 4. It is obvious that except slightly less when $\varepsilon = 0.06$, the robustness of AutoMO-Mixer is better than the other as a whole.

4 Conclusions

In this study, a new model termed as AutoMO-Mixer was developed for image guided diagnosis and therapy. In AutoMO-Mixer, sensitivity and specificity were considered as the objective functions simultaneously and a Pareto-optimal Mixer model set can be obtained in training stage. In testing stage, ERE was used to obtain safer and more robust results. The experimental results on OCT dataset showed that AutoMO-Mixer outperformed MLP-Mixer and ResNet-18 in balance, safe and robustness.

References

1. Zhang, Y., An, M.: Deep learning-and transfer learning-based super resolution reconstruction from single medical image. J. Healthc. Eng. **2017** (2017)
2. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. **19**, 221–248 (2017)
3. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems, vol. 30 (2017)
4. Huynh, E., et al.: CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. Radiother. Oncol. **120**(2), 258–266 (2016)
5. Vallières, M., Freeman, C.R., Skamene, S.R., El Naqa, I.: A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys. Med. Biol. **60**(14), 5471 (2015)
6. Błaszczyński, J., Deckert, M., Stefanowski, J., Wilk, S.: Integrating selective preprocessing of imbalanced data with Ivotes ensemble. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS (LNAI), vol. 6086, pp. 148–157. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13529-3_17
7. Chen, H., Deng, T., Du, T., Chen, B., Skibniewski, M.J., Zhang, L.: An RF and LSSVM-NSGA-II method for the multi-objective optimization of high-performance concrete durability. Cem. Concr. Compos. **129**, 104446 (2022)
8. Bagheri-Esfeh, H., Dehghan, M.R.: Multi-objective optimization of setpoint temperature of thermostats in residential buildings. Energ. Build. **261**, 111955 (2022)
9. Ayhan, M.S., Berens, P.: Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks (2018)
10. Dohopolski, M., Chen, L., Sher, D., Wang, J.: Predicting lymph node metastasis in patients with oropharyngeal cancer by using a convolutional neural network with associated epistemic and aleatoric uncertainty. Phys. Med. Biol. **65**(22), 225002 (2020)
11. Uwimana, A., Senanayake, R.: Out of distribution detection and adversarial attacks on deep neural networks for robust medical image analysis. arXiv preprint arXiv:2107.04882 (2021)
12. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)
13. Ge, Z., Wang, X.: Evaluation of various open-set medical imaging tasks with deep neural networks. arXiv preprint arXiv:2110.10888 (2021)
14. Apostolidis, K.D., Papakostas, G.A.: A survey on adversarial deep learning robustness in medical image analysis. Electronics **10**(17), 2132 (2021)
15. Paschali, M., Conjeti, S., Navarro, F., Navab, N.: Generalizability vs. robustness: investigating medical imaging networks using adversarial examples. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 493–501. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_56
16. Mangaokar, N., Pu, J., Bhattacharya, P., Reddy, C.K., Viswanath, B.: Jekyll: attacking medical image diagnostics using deep generative models. In: 2020 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 139–157. IEEE (2020)
17. Xu, M., Zhang, T., Li, Z., Liu, M., Zhang, D.: Towards evaluating the robustness of deep diagnostic models by adversarial attack. Med. Image Anal. **69**, 101977 (2021)

18. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: CutMix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)
19. Tolstikhin, I.O., et al.: MLP-mixer: an all-MLP architecture for vision. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
20. Zhou, Z., et al.: Multi-objective radiomics model for predicting distant failure in lung SBRT. Phys. Med. Biol. **62**(11), 4460 (2017)
21. Pelikan, M.: Bayesian optimization algorithm. In: Pelikan, M. (ed.) Hierarchical Bayesian Optimization Algorithm. Studies in Fuzziness and Soft Computing, vol. 170. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-32373-0_3
22. Yang, J.B., Xu, D.L.: On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **32**(3), 289–304 (2002)
23. Wang, Y.M., Yang, J.B., Xu, D.L.: Environmental impact assessment using the evidential reasoning approach. Eur. J. Oper. Res. **174**(3), 1885–1913 (2006)
24. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

Memory Transformers for Full Context and High-Resolution 3D Medical Segmentation

Loic Themyr[1,2(✉)], Clément Rambour[1], Nicolas Thome[1], Toby Collins[2], and Alexandre Hostettler[2]

[1] Conservatoire National des Arts et Métiers, Paris 75014, France
loic.themyr@lecnam.net
[2] IRCAD, Strasbourg 67000, France

Abstract. Transformer models achieve state-of-the-art results for image segmentation. However, achieving long-range attention, necessary to capture global context, with high-resolution 3D images is a fundamental challenge. This paper introduces the Full resolutIoN mEmory (FINE) transformer to overcome this issue. The core idea behind FINE is to learn memory tokens to indirectly model full range interactions while scaling well in both memory and computational costs. FINE introduces memory tokens at two levels: the first one allows full interaction between voxels within local image regions (patches), the second one allows full interactions between all regions of the 3D volume. Combined, they allow full attention over high resolution images, *e.g.* $512 \times 512 \times 256$ voxels and above. Experiments on the BCV image segmentation dataset shows better performances than state-of-the-art CNN and transformer baselines, highlighting the superiority of our full attention mechanism compared to recent transformer baselines, *e.g.* CoTr, and nnFormer.

Keywords: Transformers · 3D segmentation · Full context · High-resolution

1 Introduction

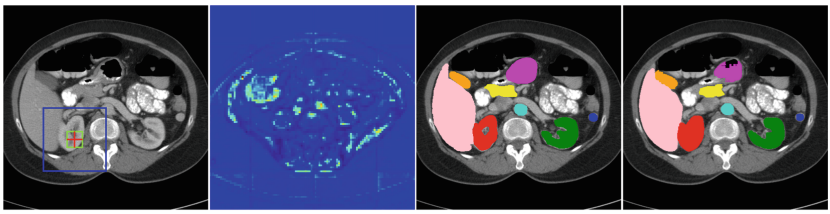
Convolutional encoder-decoder models have achieved remarkable performance for medical image segmentation [1,10]. U-Net [24] and other U-shaped architectures remain popular and competitive baselines. However, the receptive fields of these CNNs are small, both in theory and in practice [17], preventing them from exploiting global context information.

Transformers witnessed huge successes for natural language processing [4,26] and recently in vision for image classification [5]. One key challenge in 3D semantic segmentation is their scalability, since attention's complexity is quadratic with respect to the number of inputs.

Efficient attention mechanisms have been proposed, including sparse or low-rank attention matrices [21,28], kernel-based methods [12,20], window [6,16], and memory transformers [14,22]. Multi-resolution transformers [16,29,30] apply

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 121–130, 2022.
https://doi.org/10.1007/978-3-031-21014-3_13

attention in a hierarchical manner by chaining multiple window transformers. Attention at the highest resolution level is thus limited to local image sub-windows. The receptive field is gradually increased through pooling operations. Multi-resolution transformers have recently shown impressive performances for various 2D medical image segmentation tasks such as multi-organ [2,11,25], histopathological [15], skin [27], or brain [23] segmentation.



a) Input image b) FINE attention c) Ground truth d) FINE prediction

Fig. 1. Proposed full resolution memory transformer (FINE). To segment the kidney voxel in a) (red cross), FINE combines high-resolution and full contextual information, as shown in the attention map in b). This is in contrast to nnFormer [33] (resp. CoTr [31]), which receptive field is limited to the green (resp. blue) region in a). FINE thus properly segments the organs, as show in d). (Color figure online)

Recent attempts have been made to apply transformers for 3D medical image segmentation. nnFormer [33] is a 3D extension of SWIN [16] with a U-shape architecture. One limitation relates to the inherent compromise in multi-resolution, which prevents it from jointly using global context and high-resolution information. In [33], only local context is leveraged in the highest-resolution features maps. Models using deformable transformers such as CoTr [31] are able to leverage sparse global-context. A strong limitation shared by nnFormer and CoTr is that they cannot process large volumes at once and must rely on training the segmentation model on local 3D random crops. Consequently, full global contextual information is unavailable and positional encoding can be meaningless. On BCV [13], cropped patch size is about $128 \times 128 \times 64$ which only covers about 6% of the original volume.

This paper introduces the Full resolutIoN mEmory (FINE) transformer. This is, to the best of our knowledge, the first attempt at processing full-range interactions at all resolution levels with transformers for 3D medical image segmentation. To achieve this goal, memory tokens are used to indirectly enable full-range interactions between all volume elements, even when training with 3D crops. Inside each 3D crop, FINE introduces memory tokens associated to local windows. A second level of localized memory is introduced at the volume level to enable full interactions between all 3D volume patches. We show that FINE outperforms state-of-the-art CNN, transformers, and hybrid methods on the 3D multi-organ BCV dataset [13]. Figure 1 illustrates the rationale of FINE to segment the red crossed kidney voxel in a). We can see that FINE's attention map covers the whole image, enabling to model long-range interactions between

organs. In contrast, the receptive field of state-of-the art methods only cover a small portion of the volume, *e.g.* the crop size (blue) for CoTr [31] or the even smaller window's size (green) for nnFormer [33] at the highest resolution level.

2 FINE Transfomer

In this section, we detail the FINE transformer for 3D segmentation of medical images leveraging global context and full resolution information, as shown in Fig. 2. FINE is generic and can be added to most multi-resolution transformer backbones [2,16,33]. We chose to incorporate it to nnFormer [33], a strong model for 3D segmentation (see supplementary material).

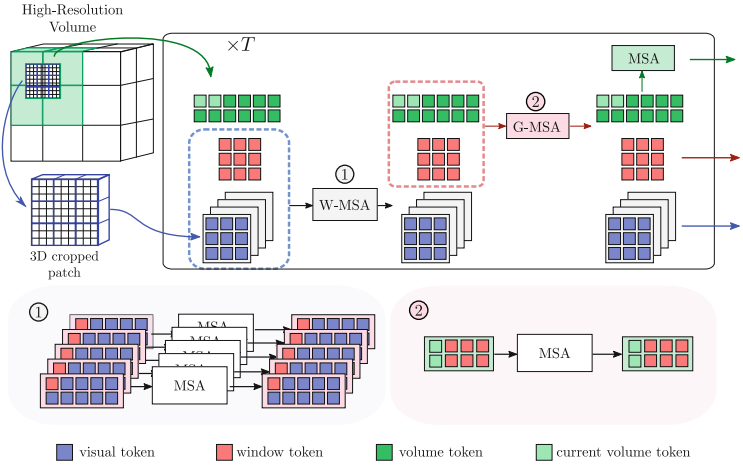


Fig. 2. To segment the cropped patch in blue and model global context, two level of memory tokens are introduced: window (red) and volume (green) tokens. First, the blue crop is divided into windows over which Multi-head Self-Attention (MSA) is performed in parallel. For each window, the sequence of visual tokens (blue) is augmented with a specific window token. Second, the local information embedded into each window token is shared between all window tokens and volume tokens intersecting with the crop (light green). Finally, high-level information is shared between all volume tokens). (Color figure online)

2.1 Memory Tokens for High Resolution Semantic Segmentation

The core idea in FINE is to introduce memory tokens to enable full-range interactions between all voxels at all resolution levels with random cropping. We introduce memory tokens at two levels.

Window Tokens. Multiple memory tokens can represent embeddings specific to regions of the feature maps [8]. Sharing these representations can thus leverage the small receptive field associated with window transformers' early stages.

In this optic, we add specific memory tokens to the sequence of visual tokens associated with each window. We chose to call them window tokens to avoid any confusion.

Volume Tokens. When dealing with high-resolution volumes, random cropping is a common training strategy. This approach is a source of limitation as only a portion of the spatial context is known by the model. Worse, no efficient positional embedding can be injected as the model has no complete knowledge of the body structure. Our memory tokens overcome this issue by keeping track of the observed part of the volume. These volume tokens are associated with each element of a grid covering the entire volume and called by the transformer blocs when performing the segmentation of a cropped patch. As can be seen Fig. 4, the volume tokens induce a positional encoding learned over the entire volume.

Discussion on Memory Tokens. The window and volume tokens can be seen as a generalisation of the class tokens used in NLP or image classification [4,5,32]. In image classification, one class token is used as a global representation of the input and sent to the classifier. In semantic segmentation, more local information needs to be preserved which requires more memory tokens.

2.2 Memory Based Global Context

Each level of memory token in FINE is related to a subdivision of the input. These memory tokens and their corresponding regions are illustrated in Fig. 2. The high-resolution volume is divided into M sub-volumes. Each sub-volume is associated to a sequence of N_w c-dimensional volume tokens $\boldsymbol{w} \in \mathbb{R}^{(M \cdot N_w) \times c}$. A 3D patch $\boldsymbol{p}$ in input of the model is divided into N windows. Each window is associated to a sequence of N_v window tokens $\boldsymbol{v} \in \mathbb{R}^{(N \cdot N_v) \times c}$. A window is composed of N_u visual tokens $\boldsymbol{u} \in \mathbb{R}^{(N \cdot N_u) \times c}$ which are the finest subdivision level.

In Fig. 2, volume, window and visual tokens are indicated in green, red, and blue respectively. First, Multi-head Self-Attention (MSA) is performed for each window over the merged sequence of visual and window tokens. Given a sequence of visual tokens *i.e.* small patches, MSA is a combination of non-local mean for all the tokens in the sequence [26]. This local operation is denoted as Window-MSA (W-MSA). Second, MSA is performed over the merged sequence of all window tokens and corresponding volume tokens to grasp long-range dependencies in the input patch. This operation is denoted as Global-MSA (G-MSA) and involve only the volume token corresponding to sub-volumes intersecting with $\boldsymbol{p}$. Finally, full resolution attention is achieved by applying MSA over the sequence of volume tokens. Formally, the t-th FINE-transformer bloc is composed of the following three operations:

$$\begin{aligned} [\boldsymbol{u}^t, \hat{\boldsymbol{v}}^t] &= \text{W-MSA}([\boldsymbol{u}^{t-1}, \boldsymbol{v}^{t-1}]), \\ [\boldsymbol{v}^t, \hat{\boldsymbol{w}}^t_{\cap}] &= \text{G-MSA}([\hat{\boldsymbol{v}}^{t-1}, \boldsymbol{w}^{t-1}_{\cap}]), \\ \boldsymbol{w}^t &= \text{MSA}(\boldsymbol{w}^t). \end{aligned} \tag{1}$$

$\boldsymbol{w}_\cap$ denotes the volume tokens corresponding to sub-volumes with a non null intersection with $\boldsymbol{p}$ and $[\boldsymbol{x}, \boldsymbol{y}]$ stands for the concatenation of $\boldsymbol{x}$ and $\boldsymbol{y}$ along the first dimension.

2.3 FINE Properties

Full Range Interactions. After two FINE transformer blocs, the memory tokens manage to capture global context in the entire volume. This global context can then be propagated to visual tokens from the current patch - see supplementary and Fig. 4.

Complexity. MSA has quadratic scaling with respect to the 3D patch dimensions while W-MSA complexity is linear with respect to the input size [16]. FINE only adds a few memory tokens and its complexity is given by:

$$\begin{aligned}\Omega(\text{FINE}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})) =& \Omega(\text{W-MSA}(\boldsymbol{u}, \boldsymbol{v})) + \Omega(\text{G-MSA}(\boldsymbol{v}, \boldsymbol{w}_\cap)) + \Omega(\text{MSA}(\boldsymbol{w})) \\ =& 2c\big(N(N_u + 2N_v) + N_{w_\cap} + MN_w\big)\big(2c + 1\big). \end{aligned} \tag{2}$$

$N_{w_\cap}$ is the number of sub-volumes intersecting with the input patch and can not exceed 8. Only a small number of global tokens brings consistent improvements and we keep $N_v = N_w = 1$. In these conditions, memory tokens are particularly efficient with a negligible complexity overhead compared to W-MSA.

3 Experiments

The Synapse Multi-Atlas Labeling Beyond the Cranial Vault (BCV) [13] dataset is used to compare performances. This dataset comprises 30 CT abdominal images with 7 manually segmented organs per image as ground truth. The organs are spleen (Sp), kidneys (Ki), gallbladder (Gb), liver (Li), stomach (St), aorta (Ao) and pancreas (Pa). The baselines are classic convolutional methods in medical image segmentation [9,18,19,24] and recent state-of-the-art transformer networks [2,3,7,31,33].

3.1 Data Preparation and FINE Implementation

All images are resampled to a same voxel spacing. The CT volumes in BCV are not centered, with strong variation along the z (cranio-caudal)-axis. To deal with this issue, the memory tokens are constant along this direction. The sub-volumes are thus reshaped with the same depth as the original volume. FINE is implemented in Pytorch and trained using a single NVidia Tesla V100-32GB GPU. All training parameters (learning rate, number of epochs, data augmentations are provided in the supplementary material). Each training epoch has 250 iterations where a randomly cropped region of size $128 \times 128 \times 64$ voxels is processed. The loss function combines multi-label Dice and cross-entropy losses, and it is optimized using SGD with a polynomial learning rate decay strategy. Deep-supervision is used during training, where the output at each decoder stage is

used to predict a downsampled segmentation mask. To avoid random noise perturbation coming from unseen memory tokens during training (typically memory tokens from regions that have never been selected), a smooth warm-up of these tokens is used. This warm-up consists of masking unseen tokens such that they do not impact the attention or the gradient.

Table 1. Method comparison using the BCV dataset and the training/test split from [33]. Average Dice scores are shown (DSC in % - higher is better). The average and individual organ 95% Hausdorff distances are also shown (HD95 in mm - lower is better). * denotes results trained by us using the authors' public code.

Method	Average		Per organ dice score (%)						
	HD95	DSC	Sp	Ki	Gb	Li	St	Ao	Pa
UNet [24]	–	77.4	86.7	73.2	69.7	93.4	75.6	89.1	54.0
AttUNet [19]	–	78.3	87.3	74.6	68.9	93.6	75.8	89.6	58.0
VNet [18]	–	67.4	80.6	78.9	51.9	87.8	57.0	75.3	40.0
Swin-UNet [2]	21.6	78.8	90.7	81.4	66.5	94.3	76.6	85.5	56.6
nnUNet [9]	10.5	87.0	91.9	86.9	**71.8**	**97.2**	85.3	**93.0**	**83.0**
TransUNet [3]	31.7	84.3	88.8	84.9	72.0	95.5	84.2	90.7	74.0
UNETR [7]	23.0	78.8	87.8	85.2	60.6	94.5	74.0	90.0	59.2
CoTr* [31]	11.1	85.7	93.4	86.7	66.8	96.6	83.0	92.6	80.6
nnFormer [33]	9.9	86.6	90.5	86.4	70.2	96.8	86.8	92.0	83.3
FINE*	**9.2**	**87.1**	**95.5**	**87.4**	66.5	97.0	**89.5**	91.3	82.5

3.2 Comparisons with State-of-the-Art

Single Fold Comparison. To fairly compare with reported SOTA results, the same single split of 18 training and 12 test images was used as detailed in [33]. The results are provided in Table 1. FINE obtains the highest average Dice score of 87.1%, which is superior to all other baselines. It also attains the best average 95% Hausdorff distances (HD95) of 9.2mm. Note that the second best method in Dice (nnUNet) is largely below FINE in HD95 (10.5), and the second best method in HD95 (nnFormer) has a large drop in Dice (86.6).

Table 2. Method comparison with SOTA transformer baselines (CoTr and nnFormer) using the BCV dataset and 5-fold cross validation. Results show mean and standard deviation of Dice (in %) for each organ and the average Dice over all organs (higher is better).

Method	Average	Sp	Ki	Gb	Li	St	Ao	Pa
CoTr [31]	84.4 ± 3.7	91.8 ± 5.0	87.9 ± 3.4	60.4 ± 10.0	95.7 ± 1.4	84.8 ± 1.3	$\mathbf{90.3} \pm 1.8$	80.0 ± 3.2
nnFormer [33]	84.6 ± 3.6	90.5 ± 6.1	87.9 ± 3.3	63.3 ± 8.1	95.7 ± 1.7	86.4 ± 0.8	89.1 ± 2.0	79.5 ± 3.5
FINE	$\mathbf{86.3} \pm 3.0$	$\mathbf{94.4} \pm 1.9$	$\mathbf{90.5} \pm 4.3$	$\mathbf{65.9} \pm 7.8$	$\mathbf{96.0} \pm 1.1$	$\mathbf{87.9} \pm 1.2$	89.4 ± 1.7	$\mathbf{80.2} \pm 2.8$
P-values	FINE vs. Cotr: 3e-2				FINE vs. nnFormer: 5e-2			

5-Fold Cross-Validation Comparison. 5-fold cross-validation of 18 training and 12 test images was used to compare FINE with the public implementation of the leading transformer baselines (CoTr and nnFormer). The Dice score results are provided in Table 2. FINE's average improvement is significant (more than 1.5 pt with the second baseline with low variance), and FINE gives the best results in 6 out of 7 organ segmentation. The statistical significance in Dice is measured with a paired 2-tailed t-test. The significance of FINE gains with respect to CoTr (3e-2) and nnFormer (5e-2) is confirmed.

3.3 FINE Analysis

Table 3. Ablation study of the impact of different tokens on BCV dataset. The metrics are Dice score (DSC in %) for all organs and in average, and the 95% Hausdorff distance (HD95 in mm). WT: Window tokens. VT: Volume tokens.

Method	WT	VT	Average		Per organ dice score						
			HD95	DSC	Sp	Ki	Gb	Li	St	Ao	Pa
nnFormer [33]	0	0	8.0	86.2	96.0	94.2	57.2	96.5	87.2	89.5	82.5
FINE	✓	0	7.7	86.6	95.7	94.2	60.9	**96.8**	85.1	90.0	**83.8**
	✓	✓	**5.2**	**87.1**	**96.2**	**94.5**	**61.5**	**96.8**	**87.3**	**90.3**	83.0

Table 4. Models memory consumption during training.

Method	Memory (GB)
nnUNet [9]	8.6
CoTr [31]	7.62
nnFormer [33]	7.73
FINE	8.05

Ablation Study. To show the impact of the different tokens in FINE, an ablation study is presented in Table 3. Three variations of FINE are compared: FINE without tokens, which is equivalent to the nnFormer method; FINE with window tokens but without volume tokens, and FINE with window and volume tokens (default). The results shows that the window tokens generally help to better segment small and difficult organs like the pancreas (Pa) and gallbladder (Gb). The use of window tokens leads to an increase in average Dice by +0.4 points. Furthermore, adding volume tokens increases performance further (average Dice increase of +0.5 points, and average HD95 reduction from 7.7 mm to 5.2 mm).

FINE Complexity. The memory consumption of FINE compared to baselines is shown in Table 4. FINE has very low overhead compared to CoTr and nnFormer. In addition, FINE has even a lower consumption than nnUNet.

Visualizations. Visualizations of segmentation results from FINE compared to CoTr and nnFormer are presented in Fig. 3. All models produce compelling results compared to the ground truth, but one can clearly see differences and especially an improved segmentation from FINE of the spleen. Visualizations of FINE attention maps are provided in Fig. 4. These attention maps show that FINE is able to leverage context information from the complete image. The left

example shows that there is attention with organs and tissues outside of the crop region (blue rectangle). Furthermore, the right example shows attention from different borders and bones like the spine, which give a strong positional information to the model.

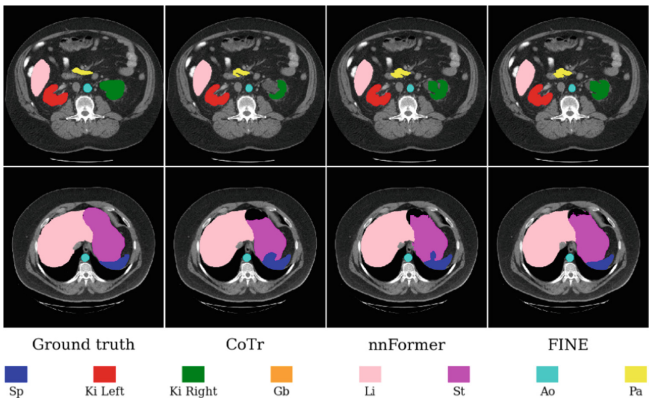


Fig. 3. Visualisation of organs segmentation by FINE compared to state-of-the-art methods on BCV. We can qualitatively see how the full context and high-resolution in FINE help in performing accurate segmentation.

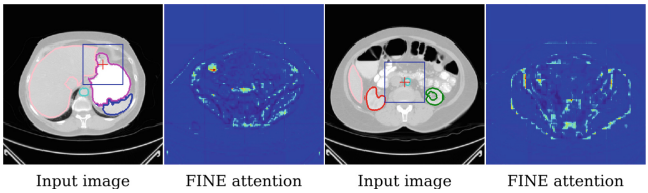


Fig. 4. Visualisation of attention maps of FINE on a BCV segmentation example. The blue rectangle is the sub-volume for which the attention has been calculated. (Color figure online)

4 Conclusion

We have presented FINE: the first transformer architecture that allows all available contextual information to be used for automatic segmentation of high-resolution 3D medical images. The technique, using two levels of memory tokens (window and volume), is applicable for any transformer architecture. Results show that FINE improves over recent and state-of-the-art transformers models. Our future work will involve the study of FINE in other modalities such as MRI or US images, as well as for other medical image tasks.

References

1. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
2. Cao, H., et al.: Swin-unet: unet-like pure transformer for medical image segmentation (2021)
3. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation (2021)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: ICLR (2021)
6. Fan, H., et al.: Multiscale vision transformers. arXiv preprint arXiv:2104.11227 (2021)
7. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation (2021)
8. Hwang, S., Heo, M., Oh, S.W., Kim, S.J.: Video instance segmentation using inter-frame communication transformers (2021)
9. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2020)
10. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017). https://doi.org/10.1016/j.media.2016.10.004. https://www.sciencedirect.com/science/article/pii/S1361841516301839
11. Karimi, D., Vasylechko, S.D., Gholipour, A.: Convolution-free medical image segmentation using transformers (2021)
12. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are RNNs: fast autoregressive transformers with linear attention (2020)
13. Landman, X., Igelsias, S., Langeral, K.: Multi-atlas labeling beyond the cranial vault. In: MICCAI (2015)
14. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: a framework for attention-based permutation-invariant neural networks. In: Proceedings of the 36th International Conference on Machine Learning, pp. 3744–3753 (2019)
15. Li, H., et al.: DT-MIL: deformable transformer for multi-instance learning on histopathological image. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 206–216. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_20
16. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: International Conference on Computer Vision (ICCV) (2021)
17. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS 2016, pp. 4905–4913. Curran Associates Inc., Red Hook (2016)
18. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation (2016)
19. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas (2018)

20. Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N., Kong, L.: Random feature attention. In: International Conference on Learning Representations (2020)
21. Qiu, J., Ma, H., Levy, O., tau Yih, S.W., Wang, S., Tang, J.: Blockwise self-attention for long document understanding (2020)
22. Rae, J.W., Potapenko, A., Jayakumar, S.M., Hillier, C., Lillicrap, T.P.: Compressive transformers for long-range sequence modelling. In: International Conference on Learning Representations (2019)
23. Reynaud, H., Vlontzos, A., Hou, B., Beqiri, A., Leeson, P., Kainz, B.: Ultrasound video transformers for cardiac ejection fraction estimation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 495–505. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_48
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
25. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: gated axial-attention for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 36–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_4
26. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
27. Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J.: Boundary-aware transformers for skin lesion segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 206–216. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_20
28. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: self-attention with linear complexity. arXiv e-prints pp. arXiv-2006 (2020)
29. Wang, W., et al.: PVTV2: improved baselines with pyramid vision transformer. arXiv preprint arXiv:2106.13797 (2021)
30. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: IEEE ICCV (2021)
31. Xie, Y., Zhang, J., Shen, C., Xia, Y.: CoTr: efficiently bridging CNN and transformer for 3D medical image segmentation (2021)
32. Zhang, P., et al.: Multi-scale vision longformer: a new vision transformer for high-resolution image encoding. In: ICCV 2021 (2021)
33. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnFormer: interleaved transformer for volumetric segmentation (2021)

Whole Mammography Diagnosis via Multi-instance Supervised Discriminative Localization and Classification

Qingxia Wu[1,2], Hongna Tan[3,4], Yaping Wu[3,4], Pei Dong[1,2], Jifei Che[5], Zheren Li[5], Chenjin Lei[5], Dinggang Shen[5], Zhong Xue[1,5(✉)], and Meiyun Wang[3,4(✉)]

[1] United Imaging Research Institute of Intelligent Imaging, Beijing, China
zhong.xue@ieee.org
[2] United Imaging Intelligence (Beijing) Co., Ltd., Beijing, China
[3] Henan Provincial People's Hospital, Henan, China
mywang@zzu.edu.cn
[4] People's Hospital of Zhengzhou University, Henan, China
[5] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China

Abstract. Precise mammography diagnosis plays a vital role in breast cancer management, especially in identifying malignancy with computer assistance. Due to high resolution, large image size, and small lesion region, it is challenging to localize lesions while classifying the whole mammography, which also renders difficulty for annotating mammography datasets and balancing tumor and normal background regions for training. To fully use local lesion information and macroscopic malignancy information, we propose a two-step mammography classification method based on multi-instance learning. In step one, a multi-task encoder-decoder architecture (mt-ConvNext-Unet) is employed for instance-level lesion localization and lesion type classification. To enhance the ability of feature extraction, we adopt ConvNext as the encoder, and added normalization layer and scSE attention blocks in the decoder to strengthen localization ability of small lesions. A classification branch is used after the encoder to jointly train lesion classification and segmentation. The instance-based outputs are merged into the image-level both for segmentation and classification (SegMap and ClsMap). In step two, a whole mammography classification model is applied for breast-level cancer diagnosis by combining the results of CC and MLO views with EfficientNet. Experimental results on the open dataset show that our method not only accurately classifies breast cancer on mammography but also highlights the suspicious regions.

Keywords: Mammography diagnosis · Multi-instance · Multi-task

Q. Wu, H. Tan and Y. Wu—Equal contribution.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 131–139, 2022.
https://doi.org/10.1007/978-3-031-21014-3_14

1 Introduction

Large-scale mammography screening programs have been successful in improving early breast cancer detection and lowering cancer-related mortality [11]. However, current human visual interpretation of mammography remains 10%–30% false-negative rate and false-positive rate [7]. Thus, it is important to improve the diagnostic performance of mammography.

Recent advances in deep learning have shown great potential in improving mammographic interpretation [14]. However, due to the high resolution of mammography (more than 4k resolution), the common computer vision models, like ResNet or DenseNet may not be suitable for mammographic images. To tackle this problem, Cai et al. [2] used a CNN to extract features within a defined region of interest (ROI), and Wang et al. [17] used 41 quantitative measurements to train a network. Both two studies achieved about 87% accuracy for micro-calcification diagnosis in a publicly available dataset [4]. Some other studies adopted the interpolation method, which resized the original image into a small size, like 264×264 [3], or 800×800 [15]. But, this method may face image deformation and uniform background problems. Beyond that, Wu et al. [17] used patches of mammograms to train a classification network first and then used patch-level classification prediction as the heatmap to train an image-level breast cancer classification. But it still faced the lack of lesion localization.

To address the aforementioned challenges, we propose a two-step breast cancer classification method to fully use local lesion information and macroscopic malignancy information. To jointly learn the lesion position information and lesion local characteristics, we propose an instance-level multi-task network mt-ConvNext-Unet to learn discriminative local lesion features. Then we merge the instance-based outputs into the image-level maps, and train an image-level network to learn the macroscopic malignancy features for mammography classification. Experimental results exhibit that our instance-level multi-task network can localize the lesion discriminatively, and further benefit the whole mammography classification.

2 Method

2.1 Method Overview

In this study, a two-step breast cancer classification method was proposed to fully use local lesion information and macroscopic malignancy information (Fig. 1). 1) An instance-level multi-task network (mt-ConvNext-Unet) was proposed to jointly learn discriminative lesion localization and lesion characteristics (normal or calcification benign or calcification malignant or mass benign or mass malignant). We randomly cropped original images into 512×512 instances and fed them into mt-ConvNext-Unet to get instance-level tumor segmentation and classification. 2) Based on the instance-level outputs, two merged image-level maps both for segmentation (SegMap) and classification (ClsMap) were created. We combined the original image with these two maps as a three-channel input

to train an EfficientNet-B0 [16] for image-level classification, aiming to learn the macroscopic malignancy features. We averaged the EfficientNet prediction from craniocaudal (CC) view and mediolateral oblique (MLO) view to get the breast-level cancer classification.

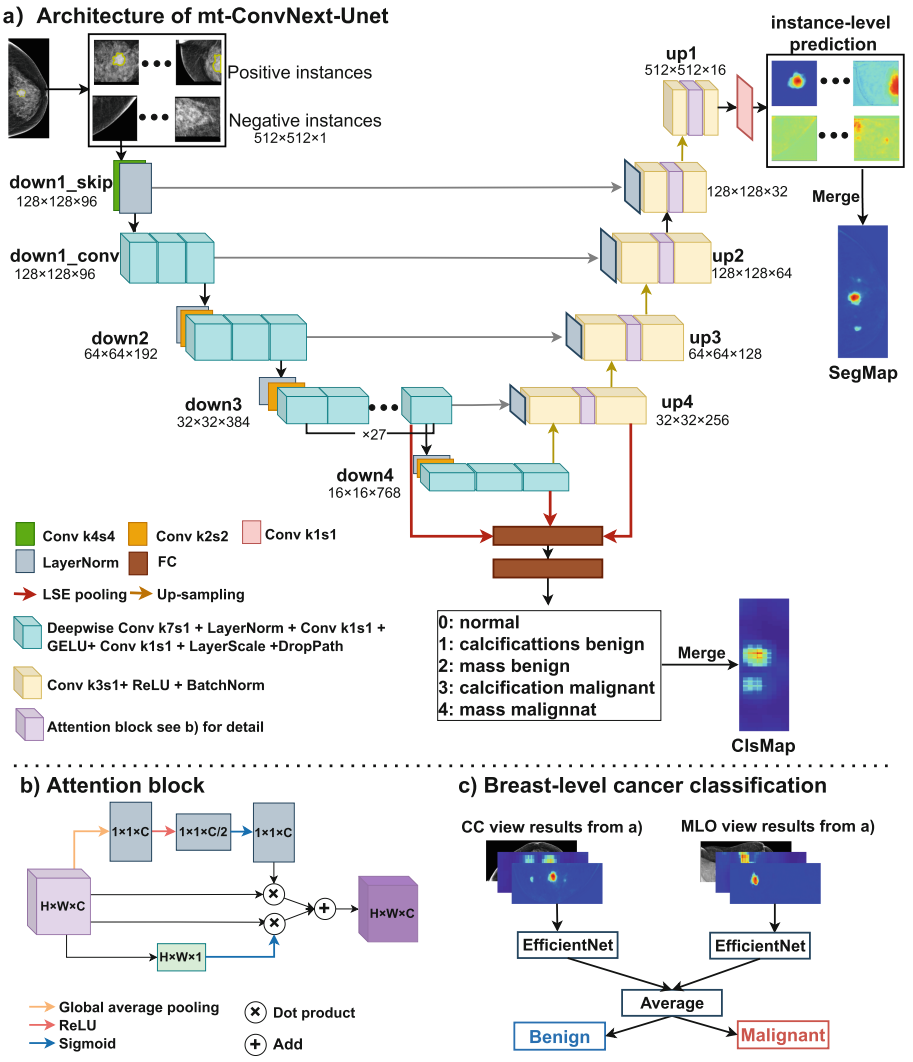


Fig. 1. Overview of the proposed method. a), step one, the architecture of instance-level multi-task learning model mt-ConvNext-Unet. b), attention block used in the decoder of mt-ConvNext-Unet. c), step two, using outputs from a) to train EfficientNet for breast-level cancer classification.

2.2 Instance-Level Multi-task Learning

We built an encoder-decoder architecture (mt-ConvNext-Unet) for instance-level multi-task learning (Fig. 1a). We used ConvNext [8] as the encoder for feature extraction due to its great performance in ImageNet top-1 accuracy and ADE20K segmentation, while we used UNet-like architecture as the decoder considering its wide application in medical image segmentation [12].

The encoder included five stages (down1_skip, down1_conv, down2, down3, and down4) and four down-sampling operations to extract different-level semantic features. The first convolutional layer in down1_skip had 128 filters with kernel size 4 and stride 4. The other down-sampling convolutional layers had kernel size 2 and stride 2. Each ConvNext block included one depthwise convolutional layer with kernel size 7 and stride 1, two convolutional layers with kernel size 1 and stride 1, one Layer Normalization (LN) [1] and Gaussian Error Linear Unit (GELU) [6].

The decoder included four up-sampling operations. We used transposed convolution layer with kernel size 2 in the first three up-sampling, and kernel size 4 in the last up-sampling. To propagate spatial information, we used skip connections to integrate the encoder features with decoder. And convolution layer with kernel size 3 following by rectified linear unit (ReLU) and batch normalization (BN) was applied in the decoder. At the end of the decoder, a convolutional layer with kernel size 1 was used to get segmentation outputs.

In addition, we added LN layers before each skip connections to normalize the different levels of encoder feature maps. We also added scSE attention [13] blocks in the decoder to strengthen the spatial and channel information, so that the network can focus on some small lesions, like micro-calcifications.

We further added a classification branch to mt-ConvNext-UNet to learn lesion characteristics, and fused feature maps from down3, down4, up4 to enhance the ability of the network to capture features for small lesions. Then we used Log-Sum-Exp (LSE) pooling before connecting different size of feature maps. The LSE pooling is more robust to image spatial information translation, and it can be formulated as:

$$x_{lse} = \frac{1}{r} \cdot \log \left[\frac{1}{H \cdot W} \cdot \sum_{i,j} \exp\left(r \cdot x_{ij}\right) \right], \tag{1}$$

where x_{lse} denotes the LSE pooling output. H and W denotes the height and width of input feature. $x_{i,j}(1 \leq i \leq H, 1 \leq j \leq W)$ denotes the pixel value at position (i, j). r is the smooth parameter, and when r goes to zero, it is average pooling; when r goes to infinite, it is max pooling. In this study, we set $r = 6$.

2.3 Losses

For the instance-level multi-task learning, we jointly trained tumor segmentation and classification. As for the segmentation task, we do not need a precise tumor boundary, so we chose Jaccard loss. In order to tackle the imbalance between

foreground and background, we combined Jaccard loss and focal loss as the segmentation loss, and they can be formulated as:

$$\mathcal{L}_{Jaccard} = 1 - \frac{(y \cdot \hat{y}) + \varepsilon}{(y + \hat{y} - y \cdot \hat{y}) + \varepsilon}, \tag{2}$$

$$\mathcal{L}_{focal} = -y\alpha(1-\hat{y})^{\gamma}\log(pr) - (1-y)\alpha\hat{y}^{\gamma}\log(1-pr), \tag{3}$$

and

$$\mathcal{L}_{seg} = \mathcal{L}_{Jaccard} + \lambda_{focal}\mathcal{L}_{focal}, \tag{4}$$

where y and $\hat{y}$ denote the ground truth and predicted probability from the instance-level segmentation. ε is a small positive number used to prevent zero denominator. α, λ_{focal} are 0.2, 5, respectively.

As for the five-category classification task in the instance-level multi-task learning, we used weighted cross entropy loss as following:

$$\mathcal{L}_{cls} = -\sum_{c}(W_c y_c \log(\hat{y}_c) + (1-y_c)\log(\hat{y}_c)), \tag{5}$$

where y_c and $\hat{y}_c$ denote the ground truth and predicted probability from the instance-level classification. W_c is the weight for each category.

The total loss for the instance-level multi-task learning is the linear combination of segmentation loss and classification loss, and it is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{cls}\mathcal{L}_{cls}, \tag{6}$$

where $\mathcal{L}_{seg}$ and $\mathcal{L}_{cls}$ denote the instance-level segmentation and classification loss, respectively, and $\mathcal{L}_{total}$ is the instance-level total loss. λ_{cls} is 0.1.

After training mt-ConvNext-Unet, we got SegMap and ClsMap for both CC and MLO view images. We combined two maps with original images in the following breast-level classification model, EfficientNet, where binary cross entropy loss was used, and it is defined as:

$$\mathcal{L}_{BCE} = -(\beta p\log(\hat{p}) + (1-\beta)(1-p)\log(1-\hat{p})), \tag{7}$$

where p and $\hat{p}$ denote the ground truth and predicted probability from the breast-level classification; β is the weight for benign and malignant breasts.

3 Experiments

3.1 Dataset

In-house Dataset. A total of 4098 mammograms from the collaborative hospitals with the IRB approvals were collected. The mammography scanners included Hologic, UIH, and Siemens. The acquisition parameters were as follow: spacing = 0.05–0.07 mm, resolution = 4604 × 5859–3328 × 4096. The tumor boundary were annotated. We randomly split the in-house dataset by a ratio of 8:2 to train and validate the performance of instance-level mt-ConvNext-Unet.

The Open Dataset (CMMD). It included 2601 breasts (1775 patients) from multi-centers. 556 breasts were biopsy-proven benign, 1316 breasts were biopsy-proven malignant, and 729 breasts were normal. Each breast had two-view paired CC and MLO images, all images were acquired on the GE system. Since CMMD does not include lesion contour, we used it to infer the instance-level mt-ConvNext-Unet. After getting SegMap and ClsMap, we used CMMD to train EfficientNet for the breast-level cancer classification. Since CMMD does not have a pre-specified test set, we used 10% of CMMD to validate the breast-level performance after considering previous studies.

3.2 Experimental Settings

The training procedure consisted of two steps. In step 1, instances of size 512×512 pixels were created from in-house dataset after the original images were cropped out background using the thresholding methods and the spacing was resampled to 0.1×0.1. Data augmentation was also used when creating the instances, including random rotations by up to 10%, resizing by up to 10%, and flipping. Preprocessing consisted of normalizing pixel values to [0, 255]. When creating positive instances, a random location within the lesion boundary was selected as the center of instances. If the resulting instance had IOU of lesion mask less than 0.1, the instance was discarded and a new instance was sampled. When creating negative instances, a random location outside the lesion boundary was selected, and the resulting instance did not had any overlap with lesion mask. The model was trained with a batch size of 24, and in each batch, the ratio of positive and negative instances is 2:9. AdamW [10] optimizer was used with a learning rate of 5×e-5, and the training epoches was 100.

After instance-level training, we used sliding window method to reconstruct instance-level segmentation outputs into SegMap. To prevent stitching artifacts, two adjacent instances had 50% overlap. Due to the uncertainty towards the border of the instance predictions, we multiplied a Gaussian map for aggregation. As for ClsMap, we used grid-wise method to merge instance-level classification outputs. Finally, we normalized the whole SegMap and ClsMap to [0, 1].

In step 2, we used a popular classification model, EfficientNet-B0 for image-level training. We concatenated original image, SegMap and ClsMap as a three-channel input, and then resized to 1100×600. AdamW optimizer was used with a learning rate of 5×e-4. Final model weights were chosen by monitoring AUC performance on the validation set.

3.3 Ablation Study

In the instance-level mt-ConvNext-Unet, we conducted the ablation experiments (Table 1). Specifically, we compared the performance with following settings: 1) basic-Unet architecture in Ronneberger et al. [12]. We also added a classification branch to the basic-Unet like mt-ConvNext-Unet. 2) We changed the encoder of basic-Unet into ConvNext. 3) We changed AdamW optimizer with fixed learning rate strategy in 2) into gradually warm up learning rate with cosine annealing

schedule strategy [5,9]. 4) We used ImageNet weight for transfer learning by summing the three channel weight into one channel weight to fit the network in 3). 5) We added LayerNorm in the skip connection of Unet decoder in 4) to normalize different levels of features. 6) To strengthen the model spatial and channel feature, we added scSE attention blocks in 5), which was the final architecture of mt-ConvNext-Unet. We used IOU and Dice to evaluate the segmentation task, and accuracy and micro-AUC to evaluate the classification task in mt-ConvNext-Unet.

The experimental results from the 820 validation mammograms showed that adopting our mt-ConvNext-Unet architecture can benefit both the segmentation task and the classification task (Table 1). The IOU improved from 0.660 to 0.757, while the accuracy improved from 0.894 to 0.925.

Table 1. The ablation study of instance-level mt-ConvNext-Unet.

Methods	IOU	Dice	Accuracy	microAUC
basic Unet	0.660	0.448	0.894	0.914
ConvNeXt encoder	0.678	0.492	0.902	0.989
ConvNeXt encoder + warmup and CosineAnnealing	0.691	0.500	0.916	0.985
ConvNext encoder + warmup and CosineAnnealing + ImageNet weight	0.739	0.608	0.924	0.989
ConvNext encoder + warmup and CosineAnnealing + ImageNet weight + LayerNorm	0.748	0.622	0.922	0.992
ConvNext encoder + warmup and CosineAnnealing + ImageNet weight + LayerNorm + Attention	0.757	0.639	0.925	0.990

3.4 Breast-Level Classification Results

After training and fine-tuning the mt-ConvNext-Unet, we inferred ABCD to get SegMap and ClsMap for both CC and MLO view mammograms. Then we trained EfficientNet for CC and MLO view separately. We averaged the prediction from CC and MLO view to get the breast-level classification. We further compared only using original images as input with using original images, SegMap and ClsMap as input, and the results is shown in Table 2. The results showed that combining CC view and MLO view can improve the model performance. Adding SegMap and ClsMap with original images can benefit the model prediction for breast cancer, AUC increased from 0.799–0.868 to 0.843–0.914. Using original images, SegMap, and ClsMap combining CC and MLO view mammograms yielded the best model with AUC = 0.914, accuracy = 0.854, precision = 0.863, and recall = 0.850.

To further understand our SegMap and ClsMap, we shows two representative breasts in Fig. 2. The first three columns show original images, SegMap and ClsMap, respectively. The fourth column (CAM_ours) shows Grad-CAM from the EfficientNet using original images, SegMap and ClsMap as inputs. The fifth column (CAM_ori) shows Grad-CAM from the EfficientNet only using original

Table 2. The ablation study of breast-level cancer classification.

Inputs	Image view	AUC	Accuracy	Precision	Recall
Original images	CC	0.827	0.773	0.813	0.721
Original images	MLO	0.799	0.785	0.808	0.759
Original images	CC + MLO	0.868	0.796	0.916	0.661
Original images, SegMap, ClsMap	CC	0.843	0.797	0.877	0.699
Original images, SegMap, ClsMap	MLO	0.892	0.831	0.887	0.767
Original images, SegMap, ClsMap	CC + MLO	0.914	0.854	0.863	0.850

images as inputs. The visualization demonstrate that the EfficientNet can learn more lesion-related features to classify breast cancer after adding SegMap and ClsMap.

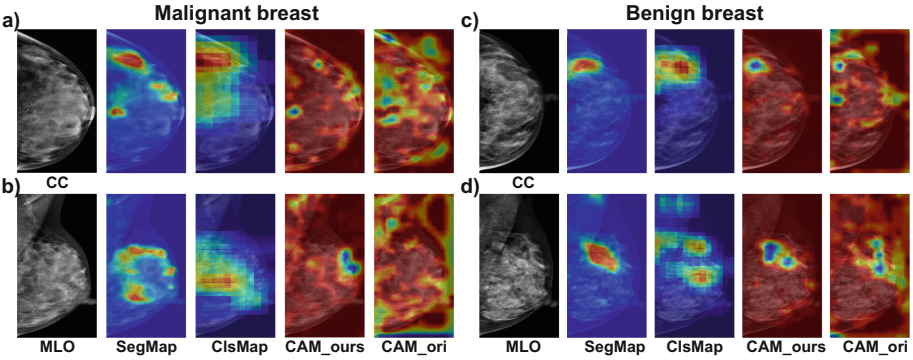


Fig. 2. Representative patients. a) and b) are malignant, c) and d) are benign.

4 Conclusion

In this paper, we proposed a two-step whole mammography classification method using instance-level discriminative lesion localization and image-level classification. Specifically, an instance-level multi-task network was employed for lesion localization and lesion type classification, and at the same time, a breast-level cancer classification network was designed to achieve macroscopic malignancy information. Different from using resizing images or placing ROIs to tackle the high-resolution mammograms in previous works, we converted multi-instance assumption into the whole mammogram attention region using SegMap and ClsMap, so as to take advantages of local instance-level lesion positioning as well as image-level characterising. Experimental results proved that our approach can localize the lesion and accurately classify breast cancer.

Acknowledgement. XZ was supported by Key R&D Program of Guangdong Province, China 2021B0101420006.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Cai, H., et al.: Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms. Comput. Math. Methods Med. **2019**, 1–10 (2019)
3. Carneiro, G., Nascimento, J., Bradley, A.P.: Automated analysis of unregistered multi-view mammograms with deep learning. IEEE Trans. Med. Imaging **36**(11), 2355–2365 (2017)
4. Chunyan, C., et al.: The Chinese mammography database (CMMD): an online mammography database with biopsy confirmed types for machine diagnosis of breast (2020). https://doi.org/10.7937/tcia.eqde-4b16
5. Goyal, P., et al.: Accurate, large minibatch SGD: training ImageNet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
6. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415 (2016)
7. Lehman, C.D., et al.: National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. Radiology **283**(1), 49–58 (2017)
8. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. arXiv preprint arXiv:2201.03545 (2022)
9. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
10. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
11. Nelson, H.D., et al.: Screening for breast cancer: an update for the US preventive services task force. Ann. Intern. Med. **151**(10), 727–737 (2009)
12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 421–429. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_48
14. Schaffter, T., et al.: Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw. Open **3**(3), e200265–e200265 (2020)
15. Shu, X., Zhang, L., Wang, Z., Lv, Q., Yi, Z.: Deep neural networks with region-based pooling structures for mammographic image classification. IEEE Trans. Med. Imaging **39**(6), 2246–2255 (2020)
16. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
17. Wu, N., et al.: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans. Med. Imaging **39**(4), 1184–1194 (2019)



Cross Task Temporal Consistency for Semi-supervised Medical Image Segmentation

Govind Jeevan, S. J. Pawan(✉), and Jeny Rajan

Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, Mangaluru, India
govindjeevan7@gmail.com, pawansnj03@gmail.com, jenyrajan@nitk.edu.in

Abstract. Semi-supervised deep learning for medical image segmentation is an intriguing area of research as far as the requirement for an adequate amount of labeled data is concerned. In this context, we propose Cross Task Temporal Consistency, a novel Semi-Supervised Learning framework that combines a self-ensembled learning strategy with cross-consistency constraints derived from the implicit perturbations between the incongruous tasks of multi-headed architectures. More specifically, the Signed Distance Map output of a teacher model is transformed to an approximate segmentation map which acts as a pseudo target for the student model. Simultaneously, the teacher's segmentation task output is utilized as the objective for the student's Signed Distance Map derived segmentation output. Our proposed framework is intuitively simple and can be plugged into existing segmentation architectures with minimal computational overhead. Our work focuses on improving the segmentation performance in very low-labeled data proportions and has demonstrated marked superiority in performance and stability over existing SSL techniques, as evidenced through extensive evaluations on two standard datasets: ACDC and LA.

Keywords: Semi supervised learning · Convolutional neural networks · Medical image segmentation

1 Introduction

Image segmentation plays an important role in extracting and quantifying the regions of interest (ROI) from various medical images. Over the past few years, convolutional neural network (CNN)-based approaches have evolved as the most successful methods for medical image segmentation. However, these methods are data-hungry and require a large number of labeled samples to build reliable models. Moreover, acquiring massive amounts of labeled data is a tedious task, as it is highly labor-intensive and time-consuming. Semi-supervised learning (SSL)

G. Jeevan and S. J. Pawan—Equal contribution.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 140–150, 2022.
https://doi.org/10.1007/978-3-031-21014-3_15

is the most practical and ideal procedure that reduces the monotonous labeling process by efficiently using the unlabeled data with a handful of labeled data to improve performance over the supervised baseline. The commonly used semi-supervised learning approaches are self-training, adversarial procedures, and consistency methods. Self-training or pseudo labeling is a naive SSL approach involving the prediction on unlabeled data $D_{UL(x_i)}$ by training the supervised baseline $D_{L(x_i,y_i)}$ and meticulous integration of unlabeled data $D_{UL(x_i)}$ with its pseudo label p_i and retraining with $(D_{L(x_i,y_i)} + D_{UL(x_i,p_i)})$ [1]. Li et al. [2] presented a generalized ensemble approach for self-training by adopting the weight sharing approach. In [3,4], uncertainty estimation is incorporated into a self-training procedure to enhance the performance. In adversarial methods, 2 sub-networks, Generator G and Discriminator D, contend with each other to achieve the objective of SSL. This could be achieved by setting up the discriminator to distinguish the prediction of labeled data $D_{L(x_i,y_i)}$ and unlabeled data $D_{UL(x_i)}$ [5] or by generating synthetic samples [6,7] by minimizing the adversarial loss L_{adv}.

Consistency regularization methods bestowed superior performance in the SSL space, compelling consistent prediction $F(D_{UL(x_i)}) = F(D_{UL(x_i)} + \delta)$ on the unlabeled data point $D_{UL(x_i)}$ subjected to data-level and task-level perturbations. In [8], Yu et al. presented an uncertainty-aware self-ensembling approach, which uses Monte Carlo-dropout minimizing the consistency loss of the student model w.r.t teacher model's output. Li et al. [9] presented a multi-task adversarial framework based on the geometric constraint by employing Signed Distance Map to make the network *shape-aware* along with the segmentation predictions. In [10], Ouali et al. presented cross-consistency training (CCT), which operates with k auxiliary decoders for unlabeled data apart from the *main-decoder* operated with the labeled data. CCT estimates the unsupervised loss by calculating the consistency between the *main-decoder* to the auxiliary decoders. A task-level regularization, namely dual-task consistency, was introduced [11] by employing a multi-head model predicting 2 inter-convertible tasks (pixel-level predictions and level-set function). Chen et al. [12] proposed cross pseudo supervision, a novel consistency regularisation method, where the same sample is subjected to two differently perturbed segmentation networks with cross-wise supervision. Several techniques using modified [13,14] architectures and semi-supervised loss functions have been presented in the literature.

This paper is structured as follows: We present the proposed methodology in Sect. 2, followed by experiments and results in Sect. 3. Finally, we conclude the paper in Sect. 4.

2 Method

2.1 Temporal Consistency Using Self-ensembling

Temporal consistency encourages the model to maintain consistency in prediction over successive iterations. Previous works such as Mean-Teacher [16], and Temporal Ensembling [15] have demonstrated the efficacy of employing temporal consistency in semi-supervised learning. The self-ensembling strategy used in

Mean-Teacher [16] involves a student and teacher ensemble, where the trainable student represents the current state of the network, and the teacher is the moving average of the past states of the network. Supervision from the teacher network's outputs encourages the student to generate consistent predictions over time.

2.2 Task Consistency with Multi Headed Architecture

In multi-headed architectures, multiple task heads are fitted onto a single shared backbone. The shared backbone is primarily responsible for extracting features from the input image, whereas the task heads use the extracted feature maps as input to predict the desired task-output. Task consistency encourages coherence between the information represented by the task outputs. Training with a task consistency framework over a shared backbone facilitates information transfer between the parallel tasks [11], enabling the learning of robust hidden representations. Additionally, it can promote better feature selection by inherently prioritizing features that carry across multiple tasks while reducing the risk of over-fitting.

2.3 Cross Task Temporal Consistency

For the proposed Cross Task Temporal Consistency (CTTC) framework, segmentation networks are constructed by fitting a U-Net [17]/V-Net [18] backbone with two prediction heads. They are separately responsible for the tasks of estimating the input's segmentation map and its signed distance map. While the model's primary objective is to predict segmentation maps, regressing over the SDMs enables the model to learn from the distinctive shapes of target ROI. To promote temporal consistency, i.e., the invariance of predictions over successive iterations, we employ a self-ensembled framework comprised of two identical multi-task networks f^{θ} and $f^{\theta'}$. f^{θ} is the student network trained using back-propagation on supervised and unsupervised losses. $f^{\theta'}$ is the teacher network with parameters θ' that are generated in every iteration from the moving average of the student network. Notably, updating the teacher network does not involve costly operations such as gradient computation and back-propagation.

Supervised Training: The segmentation prediction head of the student model, f_1^{θ} is trained using a DSC loss [19] with the segmentation masks (y) of the labeled samples (x) as shown in Eq. 1.

$$L_{DSC} = 1 - \frac{2 \sum |y \cap f_1^{\theta}(x)|}{\sum |y| + \sum |f_1^{\theta}(x)|} \tag{1}$$

In order to train the SDM prediction head of the student model (f_2^{θ}), ground truth SDMs have to be computed from segmentation labels during pre-processing. Following [11,20], we define a transformation function T (Eq. 2) that generates the SDM for a given segmentation mask y as:

$$T(i) = \begin{cases} -\inf_{j \in \partial y} \|i - j\|_2, & i \in y_{in} \\ 0, & i \in \partial y \\ \inf_{j \in \partial y} \|i - j\|_2, & i \in y_{out} \end{cases} \tag{2}$$

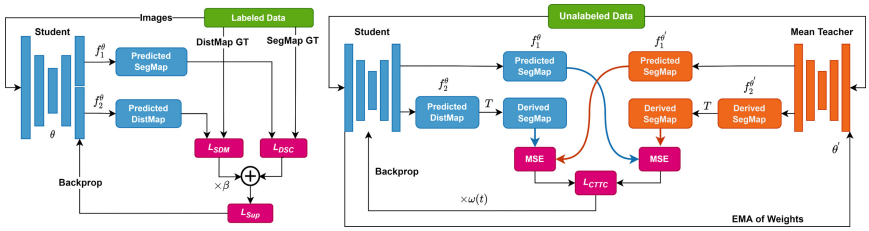


Fig. 1. The supervised (left) and unsupervised (right) training pipelines of the proposed CTTC method.

where, y_{in}, y_{out} and ∂y denote the inside, outside and boundary of target in the segmentation mask respectively. Ground truth SDMs thus generated are used to train the SDM prediction head (f_2^θ) using a consistency loss as given in Eq. 3:

$$\mathcal{L}_{\text{SDM}} = \frac{1}{|b_l|} \sum (T(y) - f_2^\theta(x))^2 \tag{3}$$

In Eq. 3, x and y are images and labels in a batch of labeled samples b_l.

Unsupervised Training: Unsupervised training is achieved by enforcing coherence between the outputs of incongruous but inter-convertible tasks on unlabeled samples as shown in Eq. 4.

$$\mathcal{L}_{\text{CTTC}} = \frac{1}{|b_{ul}|} \sum_{x \in b_{ul}} [f_1^\theta(x) - T^{-1}(f_2^{\theta'}(x))]^2 + [T^{-1}(f_2^\theta(x)) - f_1^{\theta'}(x)]^2 \tag{4}$$

To devise this consistency, the SDM predictions are converted into approximate segmentation maps by applying a differentiable Heaviside step function [11,20], denoted by T^{-1}. The approximate segmentation map derived from transforming the SDM prediction of the teacher model functions as a pseudo target for the student model's segmentation task. Similarly, the teacher model's segmentation task output is utilized as the student model's SDM task target through task-transformation [21,22].

Overall Training: The combined loss function (Eq. 5) balances the auxiliary task using a weight-coefficient β and gradually increases the priority of the unsupervised loss using an iteration dependent Gaussian warming-up function $\omega(t)$ [16]. The component losses of Eq. 5 are computed as described in the previous sections.

$$L_{overall} = L_{DSC} + \beta \cdot L_{SDM} + \omega(t) \cdot L_{CTTC} \tag{5}$$

After the student parameters θ are updated by backpropagation over $L_{overall}$, the teacher model is updated with its moving average θ'. The overall training pipeline of the student-teacher ensemble in the proposed CTTC methods is summarized in Fig. 1 and detailed in Algorithm 1.

Algorithm 1. Cross Task Temporal Consistency Algorithm.

Require:

1. $D_L(x, y)$: Collection of labeled samples
2. $D_{UL}(x)$: Collection of unlabeled samples
3. α: Rate of moving average for teacher update.
4. $\omega(t)$: Iteration dependent ramp-up function.
5. $T(y_i)\ \forall\ (x_i, y_i) \in b_l$ Pre-computed SDMs for all segmentation labels
6. DSC: Dice Coefficient function
7. MSE: Mean Squared Error function

for $t = 1, \ldots, T$ **do**
  Sample $(x_l, y_l) \sim D_L(X, Y)$
  Sample $(x_{ul}) \sim D_{UL}(X)$
  $L_{DSC} = 1 - DSC(f_1^{\theta}(x_l), y_l)$ Eq. 1
  $L_{SDM} = MSE(f_2^{\theta}(x_l), T(y_l))$ Eq. 3
  $L_{CTTC} = MSE(f_1^{\theta}(x), T^{-1}(f_2^{\theta'}(x_{ul}))) + MSE(T^{-1}(f_2^{\theta}(x_{ul})), f_1^{\theta'}(x_{ul}))$ Eq. 4
  $L_{overall} = L_{DSC} + \beta \cdot L_{SDM} + \omega(t) \cdot L_{CTTC}$ Eq. 5
  $g_\theta \leftarrow \nabla_\theta\ L_{overall}$ Compute the gradients.
  $\theta' = a\theta' + (1 - \alpha)\theta$ Update the Moving Average of the Parameters.
  $\theta \leftarrow Step(\theta, g_\theta)$ SGD
end for

3 Experiments and Results

3.1 Datasets

Left Atrial (LA) Segmentation Challenge [23]**:** The LA dataset contains 3D MRI images and segmentation labels of the left atrial cavity for 100 patients. We employ the same pre-processing techniques as Yu et al. [8] and create two groups of 80 and 20 cases for training and evaluation, respectively. Two 3D images are included in each patient case.

Automatic Cardiac Diagnosis Challenge (ACDC) [24]**:** The ACDC dataset features images from 100 patients and the corresponding multi-class segmentation labels depicting the left ventricle (LV), the myocardium (Myo), and the right ventricle (RV). The dataset is split into train, validation, and test splits with 70, 10, and 20 patient cases, respectively, with two images in each case. We follow Bai et al. [25] in processing the dataset as 2D slices rather than 3D volumes due to the sizeable inter-slice spacing in ACDC.

3.2 Experimental Setup and Implementation

The proposed method is implemented by extending an open-source SSL framework [26] for medical image segmentation. The implementations of other SSL methods used in our comparisons are also sourced from this work. In the interest of maintaining uniformity in experimental conditions and facilitating fair comparisons, we train all models from scratch on the same hardware. The training setup consists of an Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20 GHz and an NVIDIA-Tesla V100 GPU. For 3D segmentation on the LA dataset, the multi-task V-Net [18] architecture is trained with an SGD optimizer for 6000

Table 1. Quantitative of evaluations conducted on the LA dataset (3D segmentation).

Labeled	Method	DSC	JSC	95HD	ASD
4 (5%)	Supervised [18]	37.16 ± 0.23	26.63 ± 0.2	38.06 ± 1.29	12.22 ± 0.31
	DTC [11]	80.46 ± 0.79	68.55 ± 0.89	18.39 ± 0.52	4.67 ± 0.26
	ICT [27]	82.47 ± 0.34	70.66 ± 0.65	12.78 ± 0.05	2.82 ± 0.03
	MT [16]	82.81 ± 0.63	71.05 ± 0.91	13.04 ± 1.06	3.09 ± 0.16
	UAMT [8]	78.61 ± 0.05	65.62 ± 0.31	27.79 ± 3.56	8.3 ± 1.26
	EM [28]	80.42 ± 1.17	68.19 ± 1.62	25.37 ± 4.39	7.66 ± 1.28
	SASSNet [30]	79.44 ± 1.5	66.76 ± 2.21	26.52 ± 4.15	7.54 ± 1.36
	Proposed	**84.82 ± 0.63**	**73.93 ± 1.31**	**11.4 ± 0.2**	**2.64 ± 0.03**
8 (10%)	Supervised [18]	70.21 ± 4.16	56.84 ± 4.62	29.01 ± 3.79	8.48 ± 0.99
	DTC [11]	87.42 ± 0.19	77.93 ± 0.28	9.2 ± 0.32	2.24 ± 0.08
	ICT [27]	85.79 ± 0.27	75.54 ± 0.46	10.66 ± 0.34	2.21 ± 0.01
	MT [16]	86.57 ± 0.28	76.61 ± 0.41	10.55 ± 0.29	2.34 ± 0.19
	UAMT [8]	86.09 ± 0.34	76.02 ± 0.32	13.05 ± 2.94	3.6 ± 0.86
	EM [28]	85.59 ± 0.18	75.18 ± 0.28	14.8 ± 1.36	4.05 ± 0.41
	SASSNet [30]	86.25 ± 0.56	76.09 ± 0.81	18.0 ± 0.07	4.52 ± 0.10
	Proposed	**87.78 ± 0.06**	**78.4 ± 0.14**	**9.02 ± 0.16**	**2.1 ± 0.0**
16 (20%)	Supervised [18]	80.46 ± 2.25	69.47 ± 1.94	16.2 ± 1.49	4.26 ± 0.42
	DTC [11]	87.57 ± 0.82	78.23 ± 1.15	9.35 ± 0.84	2.26 ± 0.23
	ICT [27]	88.98 ± 0.13	80.34 ± 0.28	7.97 ± 0.23	2.02 ± 0.03
	MT [16]	87.45 ± 0.48	78.02 ± 0.66	8.95 ± 0.38	2.13 ± 0.07
	UAMT [8]	86.63 ± 0.19	76.72 ± 0.3	12.4 ± 2.88	3.34 ± 0.80
	EM [28]	85.64 ± 1.43	75.38 ± 1.97	12.36 ± 1.51	3.04 ± 0.43
	SaSSNet [30]	86.88 ± 0.82	77.17 ± 1.13	13.14 ± 1.21	3.51 ± 0.35
	Proposed	**89.9 ± 0.15**	**81.79 ± 0.32**	**6.52 ± 0.02**	**1.81 ± 0.0**
80 (100%)	Supervised [18]	91.41	84.24	5.47	1.63

iterations. The input is a sub-volume of size $112 \times 112 \times 80$ with a batch size of 4, each batch composed of 2 labeled and unlabeled samples each. For 2D segmentation on the ACDC dataset, the multi-task U-Net backbone [17] is trained with an SGD optimizer for 30k iterations. The slices are resized into 256×256 pixels, and the intensity of each slice is changed to [0, 1] before it's fed to the model in batches of 16. The initial learning rate of 0.01 is decayed by 0.1 every 2500 iterations. Following Li et al. [9], the weight coefficient β in Eq. 5 is set to 0.3. All experiments were repeated 3-times with a different seed for each trial. Table 1 and Table 2 report the mean and standard error of observations across the three trials.

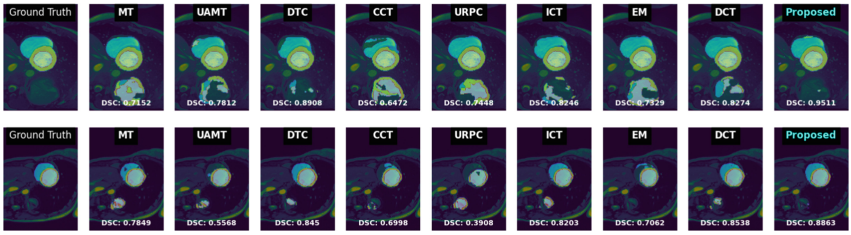


Fig. 2. The qualitative analysis on the ACDC dataset (the first column is the ground truth, and the subsequent columns are predictions of existing SSL methods, followed by the output of the proposed method).

3.3 Comparison with Other SSL Methods

We used four popular evaluation metrics; namely, Dice Similarity Coefficient (DSC), Jaccard Similarity Coefficient (JSC), Average Surface Distance (ASD), and 95% Hausdorff Distance (95HD), to quantitatively evaluate the performance of various methods considered in the study. We use the convention (D, x%) to denote a configuration where the model is trained on dataset D with x% labeled samples and the remaining treated as unlabeled samples. The results from experiments on multiple labeled data proportions are furnished in Table 1

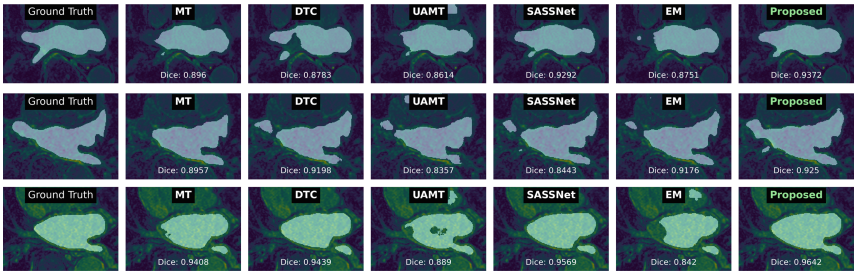


Fig. 3. The qualitative analysis on the LA dataset (the first column is the ground truth, and the subsequent columns are predictions of existing SSL methods, followed by the output of the proposed method).

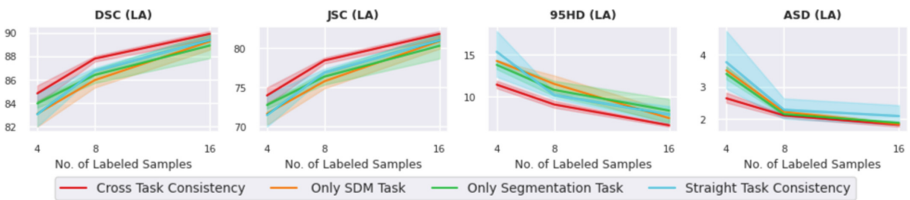


Fig. 4. The performance comparison between the straight and cross-consistency constraints for inter-convertible tasks on the LA dataset.

and Table 2. Figures 2 and 3 illustrates a qualitative analysis of the proposed method, in comparison with other SSL methods on the ACDC and LA dataset. It is evident from these observations that the proposed CTTC method demonstrates a clear superiority in performance with both LA and ACDC datasets, across all metrics. The singular case where the CTTC method stands second-best is 95HD (ACDC, 10%), where URPC has a slight edge over the proposed method. However, the same URPC model trails significantly behind CTTC and other methods in DSC and JSC metrics, implying that the strong performance trend of CTTC remains intact. Notably, CTTC (ACDC, 10%) outperforms the supervised baseline and four existing SSL methods trained with twice as much labeled data (ACDC, 20%), evidencing its proficiency in extracting meaningful information from unlabeled samples. While the proposed method offers marginal improvements in performance over other models when trained with higher proportions of labeled data, the true significance of our work is captured by the results on very low labeled data proportions. Although methods such as DTC [11] and ICT [27] produced comparable DSC with higher labeled data (LA, 20%), the performance improvement of CTTC over these methods is stark for the (LA, 5%) case, registering a gain of over 2% on DSC, and similarly for the other metrics.

Table 2. Quantitative evaluation conducted on the ACDC dataset (2D segmentation).

Labeled	Method	DSC	JSC	95HD	ASD
7 (10%)	Supervised [17]	83.69 ± 0.16	73.28 ± 0.24	6.7 ± 0.28	2.02 ± 0.07
	CCT [10]	83.69 ± 0.16	73.28 ± 0.24	6.7 ± 0.28	2.02 ± 0.07
	DTC [11]	84.97 ± 0.07	75.07 ± 0.13	9.66 ± 1.11	2.64 ± 0.29
	MT [16]	80.96 ± 1.21	69.9 ± 1.33	11.47 ± 1.4	3.2 ± 0.32
	UA-MT [8]	81.84 ± 0.66	70.86 ± 0.82	9.52 ± 0.89	2.96 ± 0.19
	URPC [28]	82.04 ± 0.38	71.14 ± 0.48	**5.47 ± 0.35**	1.7 ± 0.06
	ICT [27]	83.98 ± 0.44	73.59 ± 0.71	8.51 ± 0.1	2.58 ± 0.03
	DCT [29]	81.9 ± 0.26	71.05 ± 0.28	13.21 ± 1.42	3.54 ± 0.26
	EM [28]	82.21 ± 0.2	71.4 ± 0.25	9.22 ± 1.47	2.75 ± 0.3
	Proposed	**85.72 ± 0.01**	**75.98 ± 0.02**	6.03 ± 0.98	**1.64 ± 0.05**
14 (20%)	Supervised [17]	84.07 ± 1.15	73.81 ± 1.45	8.88 ± 0.61	2.71 ± 0.11
	CCT [10]	86.23 ± 0.25	76.92 ± 0.31	7.86 ± 0.44	2.26 ± 0.1
	DTC [11]	86.57 ± 0.31	77.67 ± 0.41	7.06 ± 1.05	2.13 ± 0.24
	MT [16]	85.14 ± 0.3	75.46 ± 0.42	9.4 ± 1.6	2.79 ± 0.37
	UA-MT [8]	85.56 ± 0.16	76.2 ± 0.3	7.01 ± 0.51	2.33 ± 0.22
	URPC [28]	85.46 ± 0.22	76.17 ± 0.32	6.04 ± 0.48	1.86 ± 0.09
	ICT [27]	85.41 ± 0.92	75.94 ± 1.11	8.15 ± 0.34	2.47 ± 0.01
	DCT [29]	84.83 ± 0.6	75.33 ± 0.66	7.03 ± 0.22	2.15 ± 0.09
	EM [28]	84.89 ± 0.2	75.15 ± 0.27	7.76 ± 0.46	2.31 ± 0.13
	Proposed	**86.71 ± 0.03**	**77.84 ± 0.08**	**5.89 ± 0.58**	**1.78 ± 0.01**
70 (100%)	Supervised [17]	91.42	84.60	2.64	0.59

Ablation Study on Consistency Strategy: We conduct additional experiments with a straight consistency constraint to strengthen the argument for the proposed cross consistency mechanism. Unlike cross-consistency, straight consistency involves supervision by the teacher over similar task outputs. The performance of the model when trained with the two tasks individually is also presented for comparison. The results visualized in Fig. 4, evidence the superiority of the proposed crossing strategy for task consistency.

4 Conclusion

Cross Task Temporal Consistency couples the stability induced by temporal consistency with the benefits of generalization and information transfer from task consistency. The inherent perturbations between the prediction outputs of incongruous tasks in multi-headed architectures are more effective than random perturbations typically used in consistency-regularization frameworks. These intuitions are substantiated through observations from extensive experiments conducted on two popular datasets, where multiple proportions of labeled data were exposed to the proposed method. CTTC outperformed existing approaches on all assessment metrics over a range of labelled data proportions, with notably large advances in extremely low labelled data proportions. The proposed method incurs minimal computational overhead and can be easily plugged into existing supervised architectures to enable learning from unlabeled samples. A potential future improvement would be to extend the cross-consistency strategy to incorporate multiple inter-transformable tasks, permuting over the set of possible consistency connections amongst them, encouraging better information transfer and reducing the risk of over-fitting further.

References

1. Lee, D.-H.: Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: ICML, pp. 03–896 (2013)
2. Li, R., Auer, D., Wagner, C., Chen, X.: A generic ensemble based deep convolutional neural network for semi-supervised medical image segmentation. In: ISBI, pp. 1168–1172 (2020)
3. Li, Y., Chen, J., Xie, X., Ma, K., Zheng, Y.: Self-loop uncertainty: a novel pseudo-label for semi-supervised medical image segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 614–623. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_60
4. Sedai, S., et al.: Uncertainty guided semi-supervised segmentation of retinal layers in OCT images. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 282–290. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_32
5. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 408–416. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_47

6. Souly, N., Spampinato, C., Shah, M.: Semi supervised semantic segmentation using generative adversarial network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5688–5696 (2017)
7. Ma, Y., et al.: Self-supervised vessel segmentation via adversarial learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7536–7545 (2021)
8. Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 605–613. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_67
9. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3D semantic segmentation for medical images. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 552–561. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_54
10. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: IEEE-CVF, pp. 12674–12684 (2020)
11. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: AAAI Conference on Artificial Intelligence (2021)
12. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: IEEE-CVF, pp. 2613–2622 (2021)
13. Lin, H., et al.: Semi-supervised NPC segmentation with uncertainty and attention guided consistency. Knowl.-Based Syst. **239**, 108021 (2021)
14. Luo, X., et al.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 318–329. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_30
15. Anneke, M., et al.: Uncertainty-aware temporal self-learning (UATS): semi-supervised learning for segmentation of prostate zones and beyond. Artif. Intell. Med. **116**, 102073 (2021)
16. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
18. Milletari, F., Navab, N., Ahmadi, S.-A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: IEEE 2016 Fourth International Conference on 3D Vision, pp. 565–571 (2016)
19. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
20. Xue, Y., et al.: Shape-aware organ segmentation by predicting signed distance maps. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12565–12572 (2020)
21. French, G., Mackiewicz, M., Fisher, M.H.: Self-ensembling for visual domain adaptation. In: International Conference on Learning Representations (2018)
22. Zhou, T., Wang, S., Bilmes, J.: Time-consistent self-supervision for semi-supervised learning. In: International Conference on Machine Learning, pp. 11523–11533 (2020)

23. Zhaohan, X., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Med. Image Anal. **34**, 101832 (2021)
24. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018). https://www.creatis.insa-lyon.fr/Challenge/acdc/index.html
25. Bai, W., et al.: Semi-supervised learning for network-based cardiac MR image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 253–260. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_29
26. Luo, X.: SSL4MIS (2020). https://github.com/HiLab-git/SSL4MIS
27. Verma, V., et al.: Interpolation consistency training for semi-supervised learning. Neural Netw. **145**, 90–106 (2022)
28. Vu, T., et al.: ADVENT: adversarial entropy minimization for domain adaptation in semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2512–2521 (2019)
29. Peng, J., et al.: Deep co-training for semi-supervised image segmentation. Pattern Recogn. **107**, 107269 (2020). ISSN 0031-3203
30. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3D semantic segmentation for medical images. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 552–561. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_54



U-Net vs Transformer: Is U-Net Outdated in Medical Image Registration?

Xi Jia[1], Joseph Bartlett[1,2], Tianyang Zhang[1], Wenqi Lu[3], Zhaowen Qiu[4], and Jinming Duan[1,5(✉)]

[1] School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK
j.duan@cs.bham.ac.uk
[2] Department of Biomedical Engineering, University of Melbourne, Melbourne, VIC 3010, Australia
[3] Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
[4] Institute of Information Computer Engineering, Northeast Forestry University, Harbin 150400, China
[5] Alan Turing Institute, London NW1 2DB, UK

Abstract. Due to their extreme long-range modeling capability, vision transformer-based networks have become increasingly popular in deformable image registration. We believe, however, that the receptive field of a 5-layer convolutional U-Net is sufficient to capture accurate deformations without needing long-range dependencies. The purpose of this study is therefore to investigate whether U-Net-based methods are outdated compared to modern transformer-based approaches when applied to medical image registration. For this, we propose a large kernel U-Net (LKU-Net) by embedding a parallel convolutional block to a vanilla U-Net in order to enhance the effective receptive field. On the public 3D IXI brain dataset for atlas-based registration, we show that the performance of the vanilla U-Net is already comparable with that of state-of-the-art transformer-based networks (such as TransMorph), and that the proposed LKU-Net outperforms TransMorph by using only 1.12% of its parameters and 10.8% of its mult-adds operations. We further evaluate LKU-Net on a MICCAI Learn2Reg 2021 challenge dataset for inter-subject registration, our LKU-Net also outperforms TransMorph on this dataset and ranks first on the public leaderboard as of the submission of this work. With only modest modifications to the vanilla U-Net, we show that U-Net can outperform transformer-based architectures on inter-subject and atlas-based 3D medical image registration. Code is available at https://github.com/xi-jia/LKU-Net.

1 Introduction

Deformable image registration, a fundamental task in medical image analysis, aims to find an optimal deformation that maps a moving image onto a fixed image. The problem can be formulated as the minimization problem including a data fidelity term that measures the distance between the fixed and warped moving image and a regularization that penalizes non-smooth deformations.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 151–160, 2022.
https://doi.org/10.1007/978-3-031-21014-3_16

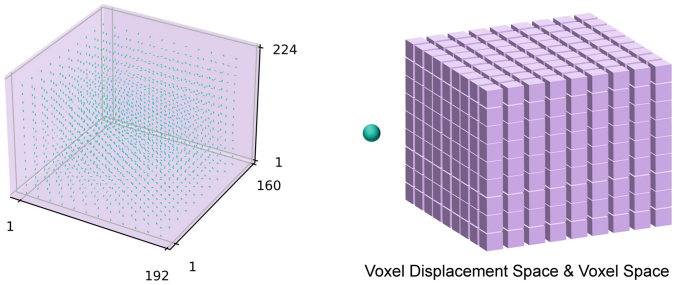


Fig. 1. Displacement fields computed from the IXI brain dataset. The left figure plots the displacement vectors in voxel averaged over the whole training data (average lengths of these vectors along x-axis, y-axis, and z-axis are 2.1 voxels, 2.3 voxels, and 1.4 voxels, respectively). The right figure is an illustration of the left figure where vectors are represented by a sphere, the size of which is much smaller than the cubic which represents the true size of the volumetric image.

Many iterative optimization approaches [3,21,24] have been proposed to tackle intensity-based deformable registration, and shown great registration accuracy. However, such methods suffer from slow inference speeds and manual tuning for each new image pair. Though some works, such as Nesterov accelerated ADMM [22], propose certain techniques to accelerate the computation, their speed still does not compare to approaches based on deep learning. Due to their fast inference speed and comparable accuracy with iterative methods, registration methods based on deep neural networks [4,5,13,18,19,25,27] have become a powerful benchmark for large-scale medical image registration.

Deep learning based registration methods [5,18,19,25] directly take moving and fixed image pairs as input, and output corresponding estimated deformations. Most deep neural networks use a U-Net style architecture [20] as their backbone and only vary preprocessing steps and loss functions. Such an architecture includes a contraction path to encode the spatial information from the input image pair, and an expansion path to decode the spatial information to compute a deformation field (or stationary velocity field). Inspired by the success of the transformer architecture [23], several recent registration works [6,7,26] have used it as the backbone to replace the standard U-Net. In this study, we investigate whether U-Net is outdated compared to modern transformer architectures, such as the new state-of-the-art TransMorph [6], for image registration.

The motivation for this work can be found in Fig. 1, where we plotted the average voxel displacement fields of the IXI brain dataset estimated by TransMorph. We notice that the average length of displacements along x-axis, y-axis, and z-axis are 2.1 voxels, 2.3 voxels, and 1.4 voxels, respectively. These displacements are very small compared to the actual volumetric size ($160 \times 192 \times 224$) of the image. Therefore, we argue that it may not be necessary to adopt a transformer to model long-range dependencies for deformable image registration. Instead, we propose a large kernel U-Net (LKU-Net) by increasing the effective receptive field of a vanilla U-Net with large kernel blocks and show that

our LKU-Net outperforms TransMorph by using only 1.12% of its parameters and 10.8% of its mult-adds operations.

2 Related Works

U-Net-Based Registration: First published in 2015, U-Net [20] and its variants have proved their efficacy in many image analysis tasks. VoxelMorph [4,5], one of the pioneering works for medical image registration, used a five-layer U-Net followed by three convolutional layers at the end. The network receives a stacked image pair of moving and fixed images and outputs their displacements. To train the network, an unsupervised loss function was used which includes a warping layer, a data term, and a regularization term. VoxelMorph has achieved comparable accuracy to state-of-the-art traditional methods (such as ANTs [3]) while operating orders of magnitude faster. Inspired by its success, many subsequent registration pipelines [13,18,25,27] used the U-Net style architecture as their registration network backbone. Among them, some works [13,27] cascaded multiple U-Nets to estimate deformations. An initial coarse deformation was first predicted and the resulting coarse deformation then refined by subsequent networks. This coarse-to-fine method usually improves the final performance, but the number of cascaded U-Nets is restricted by the GPU memory available, so training them on large-scale datasets may not be feasible with small GPUs. In this work we show that, without any cascading, a single U-Net style architecture can already outperform transformer-based networks.

Transformer-Based Registration: Transformer [23] is based on the attention mechanism, and was originally proposed for machine translation tasks. Recently, this architecture has been rapidly explored in computer vision tasks [11,15], because it successfully alleviates the inductive biases of convolutions and is capable of capturing extreme long-range dependencies. Some recent registration methods [6,7,26] have embedded the transformer as a block in the U-Net architecture to predict deformations. Building on a 5-layer U-Net, Zhang et al. [26] proposed a dual transformer network (DTN) for diffeomorphic image registration, but such a dual setting requires lots of GPU memory and greatly increases the computational complexity. As such, DTN can only include one transformer block at the bottom of the U-Net. Chen et al. [7] proposed ViT-V-Net by adopting the vision transformer (ViT) [11] block in a V-Net style convolutional network [17]. To reduce the computational cost, they input encoded image features to ViT-V-Net instead of image pairs. Their results were comparable with VoxelMorph. The authors of ViT-V-Net [7] later improved upon ViT-V-Net and proposed TransMorph [6] by adopting a more advanced transformer architecture (Swin-Transformer [15]) as its backbone. They conducted thorough experiments and showed the superiority of their TransMorph [6] over several state-of-the-art methods. In this work, we show that our proposed LKU-Net can outperform TransMorph on both inter-subject and atlas-subject brain registration tasks. We conclude in the end that fully convolutional U-Net architectures are still competitive in medical image registration.

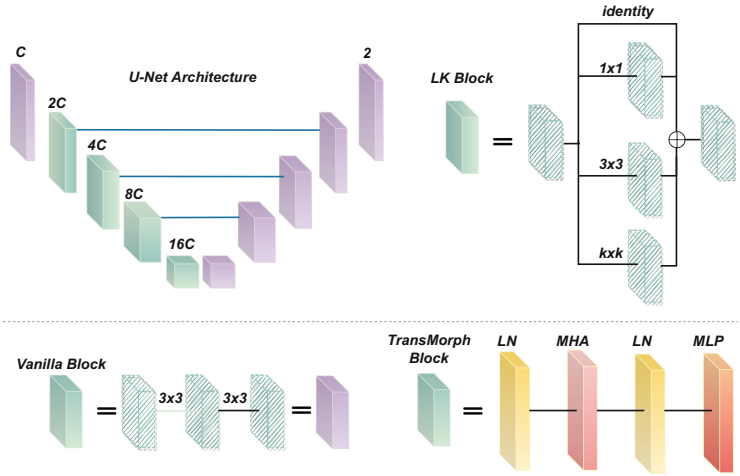


Fig. 2. Blocks used in the vanilla U-Net, TransMorph, and LKU-Net. The vanilla U-Net uses the same blocks in both the encoder and decoder, each of which consists of multiple sequential 3×3 convolutional layers. TransMorph replaces four convolutional blocks in the encoder with transformer-based blocks, each of which is built on a combination of layer norm (LN), multi-head self-attention (MHA), and multi-layer perceptron (MLP). For a fair comparison with TransMorph, we use four LK blocks in LKU-Net, each of which contains one identity shortcut and three parallel convolutional layers (that have the kernel sizes of 1×1, 3×3, and $k \times k$, respectively). The outputs of each LK block are then fused by an element-wise addition.

3 Large Kernel U-Net (LKU-Net)

Large Kernel (LK) Block: According to [1], it is easy to compute that the receptive field of a vanilla 5-layer U-Net is large enough to cover the area which could impact the deformation field around a given voxel. However, as per RepVGG [9] and RepLK-ResNet [10], in practice the effective receptive field (ERF) of a convolutional network is much smaller than the one we compute. We therefore adopt a LK block to increase the effective receptive field. Specifically, in each LK block, there are four parallel sub-layers, including a LK convolutional layer ($k \times k \times k$), a $3 \times 3 \times 3$ layer, a $1 \times 1 \times 1$ layer, and an identity shortcut. The subsequent outputs of these sub-layers are then element-wisely added to produce the output of a LK block, as shown in Fig. 2. The parallel paths in each LK block not only handle distant spatial information but also capture and fuse spatial information at a finer scale.

Directly enlarging the kernel size of a convolutional layer leads to the number of parameters growing exponentially. For example, the number of parameters increases by 463% and 1270% when enlarging a $3 \times 3 \times 3$ kernel to $5 \times 5 \times 5$ and to $7 \times 7 \times 7$, respectively. The resulting network is then cumbersome and prone to collapsing or over-fitting during training. The benefits of using the proposed LK block is that both the identity shortcut and the $1 \times 1 \times 1$ convolutional layer help the training. These numerical results are listed in our ablation studies (Table 1).

LKU-Net: We then propose LKU-Net for registration by integrating the LK blocks into the vanilla U-Net as in Fig. 2. In order to perform a fair comparison with TransMorph, which uses four transformer blocks in the contracting path of its architecture, we only use four LK blocks in the contracting path of the proposed LKU-Net. Note that, the proposed LK block is a plugin block that can be integrated into any convolutional network.

Parameterization: The resulting LKU-Net takes a stacked image pair as input, and outputs the estimated deformation. LKU-Net itself has two sets of architecture specific hyperparameters: the number of kernels and the size of the kernels in each convolutional block. For simplicity, we set the number of kernels in the first layer as C; then the number of kernels is doubled after each down-sampling layer in the contraction path and halved after each up-sampling layer in the expansion path; the number of kernels in the last layer is set to 3 for 3D displacements (2 for 2D displacements). On the other hand, though multiple LK blocks are used within our LKU-Net, we use the same kernel size $k \times k \times k$ for all LK blocks and set all other kernels to be $3 \times 3 \times 3$.

Diffeomorphism: Besides directly estimating displacements from LKU-Net, we also proposed a diffeomorphic variant, termed LKU-Net-diff, in which the final output of the LKU-Net is a stationary velocity field $\boldsymbol{v}$. We then use seven scaling and squaring layers to induce diffeomorphisms, i.e. the final deformation $\boldsymbol{\phi} = Exp(\boldsymbol{v})$ as in [2,8].

Network Loss: We adopt an unsupervised loss which consists of a normalized cross correlation (NCC) data term and a diffusion regularization term (applied to either the displacement or velocity field), which are balanced by a hyperparameter λ. The overall loss is $\mathcal{L}(\boldsymbol{\Theta})$ is $\min_{\boldsymbol{\Theta}} -\frac{1}{N}\sum_{i=1}^{N} \mathrm{NCC}(I_1^i \circ \boldsymbol{\phi}_i(\boldsymbol{\Theta}) - I_0^i) + \frac{\lambda}{N}\sum_{i=1}^{N} \|\nabla \boldsymbol{v}_i(\boldsymbol{\Theta})\|_2^2$.

Here N is the number of training pairs, I_0 denotes the fixed image, I_1 represents the moving image, $\boldsymbol{\Theta}$ are the network parameters to be learned, $\circ$ is the warping operator, and ∇ is the first order gradient implemented using the finite differences.

4 Experimental Results

Datasets: We used two datasets in our experiments. First, **OASIS** dataset [16] consists of 416 cross-sectional T1-weighted MRI scans. We used the pre-processed OASIS dataset (including 414 3D scans and 414 2D images) provided by the Learn2Reg 2021 challenge (Task 3) [12] for inter-subject brain registration. Each MRI brain scan has been skull stripped, aligned, normalized and has a resolution of $160 \times 192 \times 224$. Label masks of 35 anatomical structures were used to evaluate registration performance using metrics such as Dice Score. In this dataset, there are 394 scans (unpaired) for training, and 19 image pairs (20 scans) for validation and public leaderboard ranking. We report our 3D results on their validation set in Table 3. For fast evaluation of different methods and parameters, in Table 1,

we used 414 2D images with size 160×192, each being one slice of its respective 3D volume. We randomly split the data into: 200 images for training, 14 image pairs for validation, and 200 image pairs for testing.

Second, **IXI** dataset[1] contains nearly 600 3D MRI scans from healthy subjects, collected at three different hospitals. We used the pre-processed IXI data provided by [6]. Specifically, we used 576 T1-weighted brain MRI images to perform atlas-to-subject brain registration, in which 403, 58, and 115 images were used for training, validation, and testing, respectively. The atlas is generated by [14]. All volumes were cropped to size of $160 \times 192 \times 224$. Label maps of 29 anatomical structures were used to evaluate registration performances by Dice.

Implementation Details: The vanilla 5-layer U-Net architecture used in this work was first proposed by [25] and then used in [18], the only change we made was setting all kernels to $3 \times 3 \times 3$. 2D U-Net shares the same architecture except that all kernels are 3×3. In all experiments, we used the Adam optimizer, with batch size being set to 1, and the learning rate being kept fixed at 1×10^{-4} throughout training. Note that for the 3D OASIS registration (Learn2Reg Task 3), following [6], we additionally adopt a Dice Loss.

Ablation and Parameter Studies: In Table 1, we compare the registration performance of our LKU-Net with the vanilla U-Net using Dice on 2D OASIS data. Methods A1-A5 in Table 1 indicate that using different kernels in the vanilla U-Net affects the network's performance. Specifically, replacing all 3×3 kernels with 5×5 and 7×7 ones improves Dice by 0.09 and 0.25, respectively. However, when using 9×9 and 11×11 kernels, the performance begins to decline. Comparing the results of Methods A1 & B3, A2 & B6, A3 & C1, A4 & C2, and A5 & C3, it is easy to see that our LKU-Net outperforms the U-Net when we use the same kernels size k, and that LKU-Net is consistently better than the vanilla U-Net (A1).

Table 1. Ablation and parameter studies

Method	Model	k	C	Identity	1×1	Dice
A1	U-Net	–	8	–	–	76.16(4.08)
A2	U-Net	5	8	–	–	76.25(4.04)
A3	U-Net	7	8	–	–	76.41(4.13)
A4	U-Net	9	8	–	–	76.33(3.98)
A5	U-Net	11	8	–	–	75.80(4.03)
B1	LKU-Net	3	8	Y	N	76.26(4.18)
B2	LKU-Net	3	8	N	Y	76.40(4.06)
B3	LKU-Net	3	8	Y	Y	76.47(3.98)
B4	LKU-Net	5	8	Y	N	76.36(4.08)
B5	LKU-Net	5	8	N	Y	76.30(4.06)
B6	LKU-Net	5	8	Y	Y	76.51(4.10)
C1	LKU-Net	7	8	Y	Y	76.55(4.06)
C2	LKU-Net	9	8	Y	Y	76.45(4.03)
C3	LKU-Net	11	8	Y	Y	76.31(4.05)
D1	LKU-Net	5	16	Y	Y	77.19(3.86)
D2	LKU-Net	5	32	Y	Y	77.38(3.89)
D3	LKU-Net	7	32	Y	Y	77.52(3.90)

Meanwhile, the results from B1-B6 suggest that using either the identity shortcut or the 1×1 layer improves the registration performance, and that combining both leads to the best performance. Comparing B3, B6 and C1, we see that the performance of LKU-Net improves when we increase the kernel size. Lastly, comparing D1, D2 and D3, we find that using larger models also improves the performance, i.e., when we increase C from 8 to 16 and to 32 in LKU-Net$_{8,5}$, Dice improves from 76.51 to 77.19 and to 77.38, respectively.

[1] IXI data is available in https://brain-development.org/ixi-dataset/.

Table 2. Performance comparison between different methods on IXI. Note that the listed results (except the last six rows) are directly taken from TransMorph [6].

Model	Dice	% of $\lvert J\rvert <= 0$	Parameters	Mult-Adds (G)
Affine	0.386 ± 0.195	–	–	–
SyN	0.639 ± 0.151	<0.0001	–	–
NiftyReg	0.640 ± 0.166	<0.0001	–	–
LDDMM	0.675 ± 0.135	<0.0001	–	–
deedsBCV	0.733 ± 0.126	0.147 ± 0.050	–	–
VoxelMorph-1	0.723 ± 0.130	1.590 ± 0.339	274,387	304.05
VoxelMorph-2	0.726 ± 0.123	1.522 ± 0.336	301,411	398.81
VoxelMorph-diff	0.577 ± 0.165	<0.0001	307,878	89.67
CycleMorph	0.730 ± 0.124	1.719 ± 0.382	361,299	49.42
MIDIR	0.736 ± 0.129	<0.0001	266,387	47.05
ViT-V-Net	0.728 ± 0.124	1.609 ± 0.319	9,815,431	10.60
CoTr	0.721 ± 0.128	1.858 ± 0.314	38,684,995	1461.61
PVT	0.729 ± 0.135	1.292 ± 0.342	58,749,007	193.61
nnFormer	0.740 ± 0.134	1.595 ± 0.358	34,415,851	686.77
TransMorph	0.746 ± 0.128	1.579 ± 0.328	46,771,251	657.64
TransMorph-Bayes	0.746 ± 0.123	1.560 ± 0.333	21,205,491	657.69
TransMorph-bspl	0.752 ± 0.128	<0.0001	46,806,307	425.95
TransMorph-diff	0.599 ± 0.156	<0.0001	46,557,414	252.61
U-Net$_4$	0.727 ± 0.126	1.524 ± 0.353	279,086	58.73
U-Net-diff$_4$	0.744 ± 0.123	<0.0001	279,086	58.73
LKU-Net$_{4,5}$	0.752 ± 0.131	0.023 ± 0.018	522,302	71.00
LKU-Net-diff$_{4,5}$	0.746 ± 0.133	<0.0001	522,302	71.00
LKU-Net$_{8,5}$	0.757 ± 0.128	0.117 ± 0.058	2,086,342	272.09
LKU-Net-diff$_{8,5}$	0.753 ± 0.132	<0.0001	2,086,342	272.09

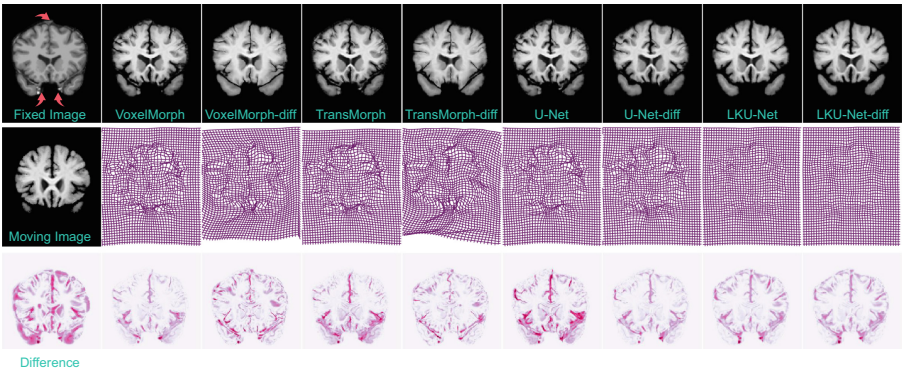


Fig. 3. Visual comparison between different methods. The first column displays the fixed image, the moving image and the difference between them. Excluding the first column, from top to bottom we show the warped moving images, the deformations, and the difference maps (between warped moving images and fixed images) of different methods. LKU-Net is able to produce smooth deformations and hence the most realistic warped moving images (see the regions marked with pink arrows). (Color figure online)

Atlas-to-Subject Registration on IXI: To guarantee a fair comparison with TransMorph[2] on this dataset, we used the exact same data pre-processing, training/validation split, and testing protocol. We report our results in Table 2, where C and k in U-Net$_C$ and LKU-Net$_{C,k}$ respectively denote the number of kernels in the first layer and the kernel size in the LK blocks.

In the case of non-diffeomorphic registration, U-Net$_4$ achieves 0.727 Dice score which is comparable with VoxelMorph-1 (0.723), VoxelMorph-2 (0.726), ViT-V-Net (0.728), CoTr (0.721), and PVT (0.729), but lower than those of nnFormer (0.740), TransMorph (0.746), and TransMorph-Bayes (0.746). However, LKU-Net$_{4,5}$ outperforms all other competing methods with a Dice score of 0.752. Note that the parameters and mult-adds of LKU-Net$_{4,5}$ are only 1.12% and 10.8% of those of TransMorph, respectively. Furthermore, by increasing the number channels from LKU-Net$_{4,5}$ to LKU-Net$_{8,5}$, a better Dice score of 0.757 can be achieved. The visual comparison of estimated deformations is given in Fig. 3. For diffeomorphic registration, U-Net-diff$_4$, LKU-Net-diff$_{4,5}$, and LKU-Net-diff$_{8,5}$ achieve Dice score of 0.744, 0.746, and 0.753, respectively. They all outperform TransMorph-diff (0.599) by a large margin and are comparable to TransMorph-bspl (0.752).

Subject-to-Subject Registration on OASIS: We further evaluated the proposed LKU-Net on the validation set of MICCAI Learn2Reg 2021 challenge (Task 3), and the results are shown in Table 3. Even though our LKU-Net contains only one single network, it has already outperformed TransMorph-LC (which is a large cascaded TransMorph). LKU-Net also outperforms LapIRN with a large margin (2.51%), which was the winner of Learn2Reg 2021 challenge. As of the submission of this work, our proposed LKU-Net holds first place on the validation leaderboard[3].

Table 3. Results of different methods on the Learn2Reg 2021 challenge Task 3.

Methods	Dice	SDlogJ	HdDist95
nnU-Net	0.8464 ± 0.0159	0.0668	1.5003
LapIRN	0.8610 ± 0.0148	0.0721	1.5139
TransMorph-LC	0.8691 ± 0.0145	0.0945	1.3969
LKU-Net	**0.8861 ± 0.0150**	0.5169	**1.2617**

5 Conclusion

With two public 3D brain datasets, we have shown that the proposed method built on a U-Net architecture can outperform modern transformer-based methods for both inter-subject and atlas-to-subject registration tasks. Our LKU-Net is conceptually simple, easy to implement, and has achieved state-of-the-art results, which proves U-Net is still competitive if devised properly. LKU-Net, however, is tested for uni-modality deformable image registration, and we will extend it to estimate deformations across different modalities.

[2] https://github.com/junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration.

[3] https://learn2reg.grand-challenge.org/evaluation/task-3-validation/leaderboard/.

Acknowledgement. The research was performed using the Baskerville Tier 2 HPC service. Baskerville was funded by the EPSRC and UKRI (EP/T022221/1 and EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham. Xi Jia is partially supported by the Chinese Scholarship Council. Joseph Bartlett is supported by the Melbourne Research Scholarship.

References

1. Araujo, A., Norris, W., Sim, J.: Computing receptive fields of convolutional neural networks. Distill **4**(11), e21 (2019)
2. Ashburner, J.: A fast diffeomorphic image registration algorithm. Neuroimage **38**(1), 95–113 (2007)
3. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage **54**(3), 2033–2044 (2011)
4. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9252–9260 (2018)
5. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
6. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: TransMorph: transformer for unsupervised medical image registration. arXiv preprint arXiv:2111.10480 (2021)
7. Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y.: Vit-V-Net: vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468 (2021)
8. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 729–738. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_82
9. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: RepVGG: making VGG-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13733–13742 (2021)
10. Ding, X., Zhang, X., Zhou, Y., Han, J., Ding, G., Sun, J.: Scaling up your kernels to 31×31: revisiting large kernel design in CNNs. arXiv preprint arXiv:2203.06717 (2022)
11. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Hoopes, A., Hoffmann, M., Fischl, B., Guttag, J., Dalca, A.V.: HyperMorph: amortized hyperparameter learning for image registration. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (eds.) IPMI 2021. LNCS, vol. 12729, pp. 3–17. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78191-0_1
13. Jia, X., et al.: Learning a model-driven variational network for deformable image registration. IEEE Trans. Med. Imaging **41**(1), 199–212 (2021)
14. Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C.: CycleMorph: cycle consistent unsupervised deformable image registration. Med. Image Anal. **71**, 102036 (2021)

15. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
16. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. **19**(9), 1498–1507 (2007)
17. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
18. Mok, T.C., Chung, A.C.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
19. Qiu, H., Qin, C., Schuh, A., Hammernik, K., Rueckert, D.: Learning diffeomorphic and modality-invariant registration using B-splines. In: Medical Imaging with Deep Learning (2021)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging **18**(8), 712–721 (1999)
22. Thorley, A., et al.: Nesterov accelerated ADMM for fast diffeomorphic image registration. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 150–160. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_15
23. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
24. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: efficient non-parametric image registration. Neuroimage **45**(1), S61–S72 (2009)
25. Zhang, J.: Inverse-consistent deep networks for unsupervised deformable image registration. arXiv preprint arXiv:1809.03443 (2018)
26. Zhang, Y., Pei, Y., Zha, H.: Learning dual transformer network for diffeomorphic registration. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 129–138. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_13
27. Zhao, S., Dong, Y., Chang, E.I.C., Xu, Y.: Recursive cascaded networks for unsupervised medical image registration. In: The IEEE International Conference on Computer Vision (ICCV) (2019)

UNet-eVAE: Iterative Refinement Using VAE Embodied Learning for Endoscopic Image Segmentation

Soumya Gupta[1,3(✉)], Sharib Ali[2], Ziang Xu[1,3], Binod Bhattarai[5], Ben Turney[6], and Jens Rittscher[1,3,4]

[1] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
soumya.gupta@eng.ox.ac.uk
[2] School of Computing, University of Leeds, Leeds, UK
[3] Big Data Institute, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Oxford, UK
[4] Oxford NIHR Biomedical Research Centre, Oxford, UK
[5] University College London, London, UK
[6] Department of Urology, The Churchill, OUH NHS Trust, Oxford, UK

Abstract. While endoscopy is routinely used for surveillance, high operator dependence demands robust automated image analysis methods. Automated segmentation of region-of-interest (ROI) that includes lesions, inflammations, and instruments can serve to cope with the operator dependence problem in this field. Most supervised methods are developed by fitting models on the available ground truth mask samples only. This work proposes a joint training approach using the UNet coupled with a variational auto-encoder (VAE) to improve endoscopic image segmentation by exploiting original samples, predicted masks and ground truth masks. In the proposed UNet-eVAE, VAE utilises the masks to constrain ROI-specific feature representations for reconstruction as an auxiliary task. The fine-grained spatial information from VAE is fused with the UNet decoder to enrich the feature representations and improve segmentation performance. Our experimental results on both colonoscopy and ureteroscopy datasets demonstrate that the proposed architecture can learn robust representations and generalise segmentation performance on unseen samples while improving the baseline.

Keywords: Endoscopy · Variational autoencoder · Segmentation · Polyp · Colonoscopy · Kidney stone · Ureteroscopy · Laser lithotripsy

1 Introduction

Endoscopy is a routine clinical procedure used to diagnose and treat hollow organs and body cavities such as the bladder, kidney, oesophagus, and colon.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_17.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 161–170, 2022.
https://doi.org/10.1007/978-3-031-21014-3_17

However, endoscopic procedures are heavily operator-dependent. The quality of endoscopy data is primarily affected by unavoidable tissue movements, illumination changes, occlusions from bodily fluids, and significant variations in organ shape and appearance [3,5]. These artefacts influence the visual interpretation of endoscopy data and may result in misdiagnosis or ineffective treatment. An example of misdiagnosis is the significantly high miss-rates for polyp detection in colonoscopy. Colonoscopy is an essential diagnostic modality for colorectal cancer (CRC) screening, with a mortality rate of 51% [7]. Another example where endoscopic artefacts may contribute to ineffective treatment is the ureteroscopic laser lithotripsy procedure for the treatment of kidney stones which affects about 12% of the world population and is associated with an increased risk of end-stage renal failure [2]. Ureteroscopic laser lithotripsy is used to fragment the stones into smaller pieces that can be removed [1]. Automated segmentation of stone and laser fibre can assist the urologists in targeting the stone accurately and reduce the excess heat generation from unnecessary lasering.

Segmentation of observed lesions in endoscopic imaging, particularly polyps, has been an active area of research. Several works are inspired by UNet-based encoder-decoder architecture [16] with some recent works extending network architecture using dual-decoders [17], double encoder-decoder networks [9], residual blocks, and attention mechanisms [8,18]. Tomar et al. [17] proposed a dual-decoders-based approach with one decoder acting as a segmentation network while the other acts as an autoencoder network that helps strengthen feature maps of the encoder. Galdran et al. [9] proposed an approach where two pretrained encoder-decoder networks are sequentially stacked, and the output from the first network acts as attention, enabling the second network to provide improved segmentation. Yeung et al. [18] proposed a dual-attention gated CNN for polyp segmentation. PraNet [8] used encoder-decoder architecture with added reverse attention mechanism for refining polyp boundaries [8]. Similarly, for the kidney stone segmentation, UNet-based architecture was used together with a CNN-based DVFNet that provided dense deformation between adjacent frames [10]. The authors showed that including such auxiliary tasks into the network improved the segmentation of kidney stones; however, the use of multiple frames to correct the segmentation makes it inefficient. Also, most of these works extended UNet by stacking modules in series [9,18] which increases network complexity and can adversely affect training on smaller datasets.

Variational auto-encoders (VAE) have been successfully used for representation disentanglement and data generation [20]. Kohl et al. proposed a generative segmentation model based on a combination of UNet with a conditional VAE to produce plausible segmentation variants for lung abnormalities segmentation [12]. Similarly, Kong et al. proposed a skip connection-based VAE-U-Net model to improve 3D brain tumour segmentation [13]. The underlying hypothesis behind this approach is that the VAE regularises the weights of the shared encoder to improve the segmentation performance. These networks attempt to project the embedding to latent space for the entire images rather than a specific region of interest. Additionally, in [12], the ground truth mask is incorporated

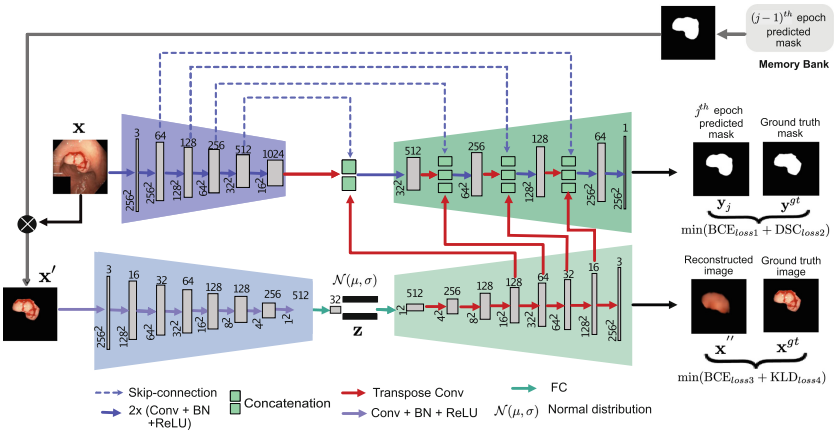


Fig. 1. Proposed UNet-eVAE framework. Here, four connections from the VAE-decoder are fused with the segmentation-decoder. Since the initial decoder layers have small spatial details $(1^2, 4^2, 8^2)$, latter layers are chosen $(16^2, 32^2, 64^2, 128^2)$.

together with original image samples. However, this will constrain VAE representations towards majority samples in training and can affect the generalisability.

To tackle such a problem and increase the bandwidth of the representations that the VAE can learn to boost segmentation, we take two major steps - 1) we aim to reconstruct the region-of-interest after applying the predicted mask from the previous epoch and aim to minimise loss with the ground truth region-of-interest (ROI) area. 2) we enrich the features of the UNet decoder layers by concatenating the features from various decoder layers of the VAE. Here, the former strategy helps to learn unseen representations during input and generalise to hard samples iteratively. At the same time, the latter strengthens the feature representations to amplify features in the regions that contribute to the segmentation map generation. Thus, in this work, we introduce a novel "UNet-eVAE" architecture that combines VAE and UNet and propose an iterative mechanism that efficiently enhances the segmentation performance. Here, we use VAE for performing auxiliary reconstruction tasks for learning ROI-specific representations and provide a feedback mechanism from our main segmentation task. We provide extensive experiments to demonstrate the performance improvement on test datasets of the same distribution and generalisability on out-of-sample datasets for colonoscopy and ureteroscopy datasets.

2 Method

The proposed framework, as shown in Fig. 1 is a novel integration of two subnetworks: UNet and VAE ("UNet-eVAE"), such that the VAE enhances ROI-specific feature representations as an auxiliary task, and the fine-grained spatial information from the VAE is fused with the UNet decoder. This technique enriches the feature representations and improves segmentation performance.

2.1 UNet Module with VAE

The UNet is an encoder-decoder architecture. The input image $\mathbf{x} \in \mathbb{R}^{3\times h\times w}$ is downsampled at consecutive encoder layers while learning feature representations through a series of convolution, batch normalisation and activation function. The decoder is then used to upsample these features and predict segmentation map $\mathbf{y}_j \in \mathbb{R}^{h\times w}$ that corresponds to each pixel classification in $\mathbf{x}$ where j refers to j^{th} epoch. The skip connections (represented by dotted lines in Fig. 1) help to recover spatial information lost during downsampling. Our proposed approach follows a joint-optimisation strategy where image $\mathbf{x}$ is fed into the UNet model to obtain a segmentation prediction mask $\mathbf{y}_j$ for j^{th} epoch. We use a variational autoencoder (VAE) to perform reconstruction of the region-of-interest area obtained from UNet architecture. For this, the segmentation mask obtained from the previous epoch $\mathbf{y}_{j-1}^{th}$ is multiplied by the original image $\mathbf{x}$ to obtain $\mathbf{x}^{'}$ $(= \mathbf{x} \odot \mathbf{y}_{j-1})$ as shown in Fig. 1. Thus, a masked input $\mathbf{x}^{'}$ is compared with the ground truth masked input $\mathbf{x}^{gt}$ $(= \mathbf{x} \odot \mathbf{y}^{gt})$ to constrain VAE reconstruction. The presented auxiliary task enriches feature representations used to obtain segmentation masks by further concatenating them with the UNet decoder layers (illustrated using red arrows from the VAE decoder to the UNet decoder in Fig. 1). Let $q_\theta(\mathbf{z}|\mathbf{x}^{'})$ be the latent space distribution of our sample $\mathbf{x}^{'}$. $\mathbf{z}$ (with parameters $\mu_\mathbf{z}$ and $\sigma_\mathbf{z}$) sampled from this distribution is fed into the decoder, $p_\phi(\mathbf{x}^{gt}|\mathbf{z})$. The loss function of the VAE is given by:

$$\mathcal{L}(\theta, \phi) = -\mathbb{E}_{q_\theta}[\ln(p_\phi(\mathbf{x}^{gt}|\mathbf{z}))] + \mathbb{KL}(q_\theta(\mathbf{z}|\mathbf{x}^{'})||p(\mathbf{z})) \tag{1}$$

2.2 Joint Loss Minimisation

Our proposed model aims to jointly optimise loss functions responsible for the main segmentation task and an auxiliary reconstruction task. For the segmentation task, we have used a combination of binary cross-entropy (BCE) BCE_{loss1} and dice similarity coefficient DSC_{loss2} [14] losses that are calculated between the predicted mask $\mathbf{y}_j$ and its ground truth mask $\mathbf{y}^{gt}$. Our proposed loss for segmentation $\mathcal{L}_{seg}$ is given by:

$$\mathcal{L}_{seg} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{y}_i^{gt}(log(\mathbf{y}_{j_i})) + (1-\mathbf{y}_i^{gt})(log(1-\mathbf{y}_{j_i})))}_{\text{BCE}_{\text{loss1}}} + \underbrace{1 - \frac{2 * \sum_{i=1}^{N}(\mathbf{y}_i^{gt} * \mathbf{y}_{j_i})}{\sum_{i=1}^{N}\mathbf{y}_i^{gt} + \sum_{i=1}^{N}\mathbf{y}_{j_i}}}_{\text{DSC}_{\text{loss2}}}, \tag{2}$$

where N is the total number of pixels, $\mathbf{y}^{gt}$ is the label (1 or 0) and $\mathbf{y}^{j}$ is the predicted probability. For the auxiliary reconstruction task, a combination of BCE reconstruction loss BCE_{loss3} and KL-divergence KLD_{loss4} is used and denoted by $\mathcal{L}_r$. Here, BCE_{loss3} is calculated between the reconstructed image from VAE network $\mathbf{x}''$ and the ground truth image $\mathbf{x}^{gt}$. $\mathcal{L}_r$ is given by:

$$\mathcal{L}_r = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^{gt}(log(\mathbf{x}_i^{''})) + (1-\mathbf{x}_i^{gt})(log(1-\mathbf{x}_i^{''})))}_{\text{BCE}_{\text{loss3}}} - \underbrace{0.5\beta\sum_{k=1}^{\mathbf{z}}(1+\log\sigma_k^2-\sigma_k^2-\mu_k^2)}_{\text{KLD}_{loss4}}, \tag{3}$$

where $\beta = 10$ (see ablation study in Table 1 **Supplementary material**) is used to weight KLD_{loss4}. The final joint loss $0.5\times(\mathcal{L}_{seg}+\mathcal{L}_r)$ is optimised using Adam optimiser with learning rate of $1e^{-4}$.

3 Experiments and Results

Implementation Details and Setup. All experiments are done in the PyTorch framework and trained on NVIDIA Quadro RTX 6000. All image samples were resized to $256 \times 256 \times 3$ pixels. Adam optimiser with a learning rate of $1e^{-4}$ and batch sizes of 4 and 32 were used, respectively, for the Ureteroscopy dataset and Kvasir-SEG dataset. ReduceLROnPlateau callback was used to monitor the learning rate and tune it to obtain optimal performance. The EarlyStopping was also used to monitor validation loss and stop the training when no improvement was observed. For our UNet-eVAE, the previous epoch segmentation masks $\mathbf{y}_{j-1}$ were initialised using Otsu thresholding [15] of original image $\mathbf{x}$ during the first training epoch, and the same was done for the model inference. Standard medical image segmentation metrics are used, such as mean intersection over union (mIoU), mean Sørensen-dice coefficient (DSC), precision, and recall.

Datasets. Both publicly available and in-house datasets are used. Publicly available Kvasir-SEG [11] was used for polyp segmentation task. The Kvasir-SEG dataset consists of 1000 polyp images that were randomly split into a training set (80%), validation set (10%), and testing set (10%) with a random seed at 42. To assess the generalisability, we use two other publicly available datasets that include CVC-ClinicDB [6] and PolypGen-C6 [4] for out-of-sample (unseen distribution) tests. CVC-ClinicDB consists of 66 samples, and PolypGen-C6 consists of 88 samples. For the ureteroscopic laser lithotripsy, *in vitro* and *in vivo* datasets that were collected in-house are used. The *in vitro* dataset consists of 70 training samples, 9 validation samples, and 9 test samples. The *in vivo* dataset consists of 136 clinically acquired images. This dataset was split into a training set (80%), validation set (10%), and testing set (10%) with a random seed at 42. To assess generalisability, we have additionally used 18 samples from an animal study treated as an out-of-sample test dataset.

3.1 Quantitative Results

Table 1 presents quantitative results on various test sets wherein dataset 1 represents Kvasir-SEG, while dataset 2 and dataset 3 represent ureteroscopy test

Table 1. Quantitative evaluation of segmentation from UNet and proposed UNet-eVAE (ours) trained on Kvasir-SEG, *in vitro* and *in vivo* ureteroscopic datasets. In the *in vivo* dataset, cl and an refer to clinical and animal samples, respectively.

Network	Class	Train set	Test set	DSC	mIOU	Precision	Recall
Dataset 1 - Colonoscopy for polyp segmentation: Trained on Kvasir-SEG							
Prob.UNet [12]	Polyp	Kvasir-SEG (450)	Kvasir-test	0.583 ± 0.284	0.457 ± 0.263	0.774 ± 0.268	0.578 ± 0.315
DDANet [17]	Polyp	Kvasir-SEG (450)	Kvasir-test	0.658 ± 0.297	0.555 ± 0.300	0.765 ± 0.306	0.649 ± 0.322
UNet [16]	Polyp	Kvasir-SEG (450)	Kvasir-test	0.677 ± 0.287	0.572 ± 0.281	**0.771** ± 0.258	0.724 ± 0.311
			CVC-ClinicDB	0.369 ± 0.331	0.283 ± 0.283	**0.557** ± 0.382	0.543 ± 0.436
			PolypGen-C6	0.369 ± 0.343	0.287 ± 0.295	**0.490** ± 0.394	0.538 ± 0.424
Ours	Polyp	Kvasir-SEG (450)	Kvasir-test	**0.726** ± 0.270	**0.628** ± 0.276	0.754 ± 0.285	**0.792** ± 0.268
			CVC-ClinicDB	**0.453** ± 0.315	**0.350** ± 0.284	0.425 ± 0.337	**0.692** ± 0.386
			PolypGen-C6	**0.399** ± 0.339	**0.313** ± 0.307	0.405 ± 0.371	**0.666** ± 0.395
Prob.UNet [12]	Polyp	Kvasir-SEG (900)	Kvasir-test	0.588 ± 0.287	0.467 ± 0.278	0.792 ± 0.261	0.632 ± 0.333
DDANet [17]	Polyp	Kvasir-SEG (900)	Kvasir-test	0.764 ± 0.252	0.671 ± 0.264	0.770 ± 0.266	**0.830** ± 0.255
UNet [16]	Polyp	Kvasir-SEG (900)	Kvasir-test	0.773 ± 0.277	0.693 ± 0.285	**0.849** ± 0.258	0.788 ± 0.277
			CVC-ClinicDB	0.472 ± 0.373	**0.390** ± 0.338	**0.749** ± 0.332	0.525 ± 0.419
			PolypGen-C6	**0.478** ± 0.397	**0.410** ± 0.375	**0.750** ± 0.348	0.528 ± 0.421
Ours	Polyp	Kvasir-SEG (900)	Kvasir-test	**0.786** ± 0.225	**0.693** ± 0.250	0.822 ± 0.247	0.824 ± 0.224
			CVC-ClinicDB	**0.488** ± 0.305	0.380 ± 0.290	0.501 ± 0.357	**0.681** ± 0.366
			PolypGen-C6	0.458 ± 0.360	0.375 ± 0.338	0.474 ± 0.385	**0.659** ± 0.401
Dataset 2 - Ureteroscopic laser lithotripsy (In-vitro)							
UNet [16]	Stone	*In vitro*	*In vitro*	0.928 ± 0.028	0.867 ± 0.048	0.920 ± 0.068	0.941 ± 0.034
Ours	Stone	*In vitro*	*In vitro*	**0.944** ± 0.031	**0.896** ± 0.054	**0.947** ± 0.053	**0.944** ± 0.029
UNet [16]	Laser	*In vitro*	*In vitro*	0.731 ± 0.305	0.647 ± 0.304	**0.770** ± 0.332	0.825 ± 0.173
Ours	Laser	*In vitro*	*In vitro*	**0.767** ± 0.312	**0.700** ± 0.310	0.764 ± 0.306	**0.884** ± 0.176
Dataset 3 - Ureteroscopic laser lithotripsy (In-vivo): Trained on clinical							
UNet [16]	Stone	*In vivo* (cl)	*In vivo*(cl)	0.886 ± 0.075	0.803 ± 0.110	0.915 ± 0.057	0.876 ± 0.131
			In vivo(an)	0.422 ± 0.327	0.324 ± 0.273	**0.928** ± 0.105	0.354 ± 0.303
Ours	Stone	*In vivo*(cl)	*In vivo* (cl)	**0.900** ± 0.061	**0.824** ± 0.091	**0.929** ± 0.037	**0.883** ± 0.111
			In vivo(an)	**0.446** ± 0.304	**0.335** ± 0.252	0.839 ± 0.244	**0.368** ± 0.280
UNet [16]	Laser	*In vivo*(cl)	*In vivo*(cl)	**0.936** ± 0.032	**0.881** ± 0.055	**0.947** ± 0.043	0.930 ± 0.066
Ours	Laser	*In vivo*(cl)	*In vivo*(cl)	0.931 ± 0.036	0.873 ± 0.062	0.913 ± 0.060	**0.954** ± 0.040

samples with inference from checkpoints that were trained on *in-vitro* and *in-vivo*, respectively. For dataset 1, results of training on 50% samples and all 900 samples are presented and compared with our proposed model and SOTA approaches such as Probabilistic UNet [12], DDANet [17] and UNet [16]. It can be observed that when trained with 450 samples only, our method provided a DSC improvement of about 24.5%, 10.4%, and 7.26%, on the Kvasir-SEG test set(100 samples) as opposed to Probabilistic UNet, DDANet and UNet, respectively. It can also be seen that our proposed approach showed improved generalisability with a 23%, and 8.13% DSC improvement on out-of-sample CVC-ClinicDB (66 samples) and PolypGen-C6 (88 samples), respectively as opposed to UNet. Similarly, in the case of 900 samples, our proposed method provided a DSC improvement of 1.7% and 3.47% on the Kvasir-SEG test set (100 samples) and CVC-ClinicDB (66 samples) respectively. It can also be observed that in the case of 900 samples, our proposed approach provided a DSC improvement of 33.8% and 2.9% over Probabilistic UNet and DDANet, respectively on the Kvasir-SEG test set. It can also be seen in this section that the segmentation results of our

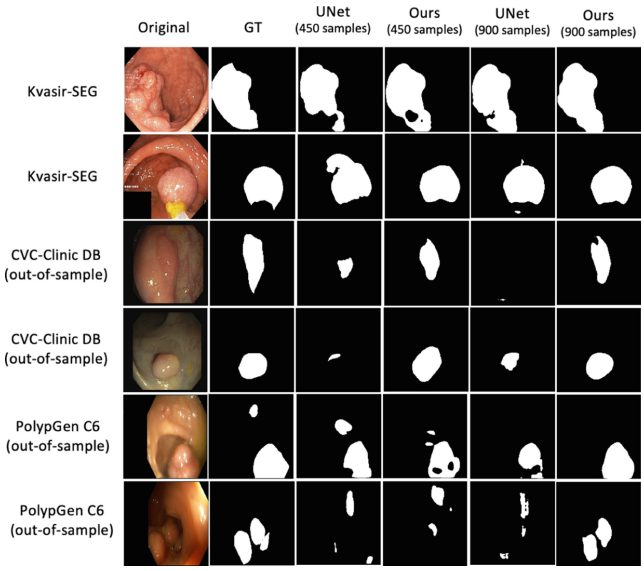


Fig. 2. Qualitative evaluation of segmentation results obtained from UNet and UNet-eVAE trained on Kvasir-SEG dataset

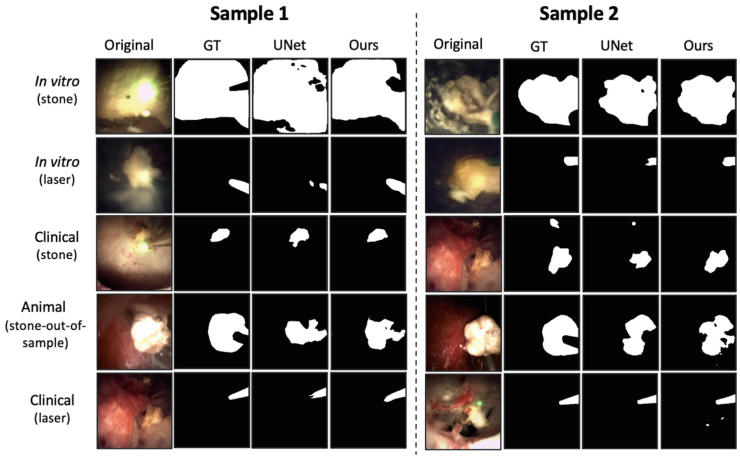


Fig. 3. Qualitative evaluation of segmentation results obtained from UNet and UNet-eVAE trained on *in vitro* and *in vivo* ureteroscopic laser lithotripsy datasets.

proposed method have a lower standard deviation across all datasets for polyp segmentation as compared to the results obtained from SOTA. It can also be observed that the performance of existing SOTA approaches significantly lowers when the training set is halved while the proposed method still performs comparatively well. Similarly, for Dataset-2 in Table 1, it can be observed that our proposed method boosted the DSC of both stone and laser classes by 1.76% and

4.98%, respectively. Finally, in the third section of Table 1 (refer to Dataset 3) for which results of models trained using 136 *in vivo* (clinical) samples, it can be seen that for the stone class, our proposed method provided a DSC improvement of 1.6% on the clinical test set (14 samples) and also improved the generalisation on out-of-sample animal data (18 samples) by 5.53% as opposed to SOTA UNet. For the laser class, our model provided competitive results to UNet.

3.2 Qualitative Results

In Fig. 2, the first two rows represent the network inference results on Kvasir-SEG, while the next four rows illustrate the network inference on the out-of-sample CVC-ClinicDB and PolypGen-C6 datasets. Our proposed model significantly outperforms the UNet model on the Kvasir-SEG test set and generalises well on both out-of-sample datasets compared to the UNet model. It can also be seen that the performance of our proposed model is better than the UNet even when the number of training samples is reduced to half. Figure 3 illustrates results on ureteroscopic datasets for two samples for each provided case in the rows. The first two rows show segmentation results of stone and laser classes in the *in vitro* dataset. It can be observed that our proposed method yields closer to ground truth masks compared to the UNet model. Similar results can be observed in the third and fourth rows for stone class in the clinical and animal datasets (out-of-sample), respectively. UNet-eVAE provided competitive segmentation performance on clinical samples even for the laser class.

4 Discussion and Conclusion

We have presented a novel end-to-end trainable method that uses a VAE as an auxiliary network to enhance ROI-specific feature representations that are combined and optimised with a UNet for endoscopic image segmentation tasks. Our proposed method outperforms SOTA UNet and shared-encoder-based approaches such as Probabilistic UNet [12] and DDANet [17]. It uses the previous epoch's prediction to constrain the VAE only to provide robust fine-grained ROI-specific features that act as attention to the segmentation network, thereby providing improved segmentation. The added strength of the proposed method is its generalisability and that it can boost the performance of any baseline network independent of its architecture (e.g. DeepResUNet [19] yielded a 14% boost with our technique - see **Table 2 Supplementary material**). Our experimental results on varied endoscopy datasets demonstrate the ability of the proposed UNet-eVAE to generalise on both same-distribution test samples and unseen out-of-sample test datasets.

Acknowledgement. We would like to thank Boston Scientific for funding this project (Grant No: DFR04690). SG and BT are funded by BSC, BB is funded by EndoMapper Horizon 2020 FET (GA 863146), SA and JR were supported by the NIHR Oxford Biomedical Research Centre.

References

1. Aldoukhi, A.H., Roberts, W.W., Hall, T.L., Ghani, K.R.: Holmium laser lithotripsy in the new stone age: dust or bust? Front. Surg. **4**, 57 (2017)
2. Alelign, T., Petros, B.: Kidney stone disease: an update on current concepts. Adv. Urol. **2018** (2018)
3. Ali, S., et al.: Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Med. Image Anal. **70**, 102002 (2021)
4. Ali, S., et al.: PolypGen: a multi-center polyp detection and segmentation dataset for generalisability assessment. arXiv preprint arXiv:2106.04463 (2021)
5. Ali, S., et al.: An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Sci. Rep. **10**(1), 1–15 (2020)
6. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. **43**, 99–111 (2015)
7. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **68**(6), 394–424 (2018)
8. Fan, D.-P., et al.: PraNet: parallel reverse attention network for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 263–273. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_26
9. Galdran, A., Carneiro, G., Ballester, M.A.G.: Double encoder-decoder networks for gastrointestinal polyp segmentation. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12661, pp. 293–307. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68763-2_22
10. Gupta, S., Ali, S., Goldsmith, L., Turney, B., Rittscher, J.: MI-UNet: improved segmentation in ureteroscopy. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 212–216 (2020). https://doi.org/10.1109/ISBI45749.2020.9098608
11. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
12. Kohl, S., et al.: A probabilistic U-Net for segmentation of ambiguous images. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
13. Li, K., Kong, L., Zhang, Y.: 3D U-Net brain tumor segmentation using VAE skip connection. In: 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC), pp. 97–101. IEEE (2020)
14. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
15. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Tomar, N.K., et al.: DDANet: dual decoder attention network for automatic polyp segmentation. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12668, pp. 307–314. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68793-9_23

18. Yeung, M., Sala, E., Schönlieb, C.B., Rundo, L.: Focus U-Net: a novel dual attention-gated CNN for polyp segmentation during colonoscopy. Comput. Biol. Med. **137**, 104815 (2021)
19. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geosci. Remote Sens. Lett. **15**(5), 749–753 (2018)
20. Zhu, Y., Min, M.R., Kadav, A., Graf, H.P.: S3VAE: self-supervised sequential VAE for representation disentanglement and data generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6538–6547 (2020)



Dynamic Linear Transformer for 3D Biomedical Image Segmentation

Zheyuan Zhang and Ulas Bagci(✉)

Northwestern University, Evanston, IL 60201, USA
ulas.bagci@northwestern.edu

Abstract. Transformer-based neural networks have surpassed promising performance on many biomedical image segmentation tasks due to a better global information modeling from the self-attention mechanism. However, most methods are still designed for 2D medical images while ignoring the essential 3D volume information. The main challenge for 3D Transformer-based segmentation methods is the quadratic complexity introduced by the self-attention mechanism [17]. In this paper, we are addressing these two research gaps, lack of 3D methods and computational complexity in Transformers, by proposing a novel Transformer architecture that has an encoder-decoder style architecture with linear complexity. Furthermore, we newly introduce a dynamic token concept to further reduce the token numbers for self-attention calculation. Taking advantage of the global information modeling, we provide uncertainty maps from different hierarchy stages. We evaluate this method on multiple challenging CT pancreas segmentation datasets. Our results show that our novel 3D Transformer-based segmentor could provide promising highly feasible segmentation performance and accurate uncertainty quantification using single annotation. Code is available https://github.com/freshman97/LinTransUNet.

Keywords: Linear transformer · Pancreas segmentation · Uncertainty quantification

1 Introduction

Long-range dependence is needed and essential for the challenging segmentation problems in biomedical images. Transformer is a novel architecture that allows us to capture such information and has achieved promising performance on many vision tasks including biomedical image segmentation [2,6,17]. A large amount of Transformer-based biomedical image segmentation methods has been proposed in the last two years. For example, the Swin-Unet replaces the traditional convolution blocking with the Transformer as the feature extraction and reconstruction tools in the encoder and decoder process [2]. However, it is still challenging to apply such architecture to 3D images due to the huge computation challenges. The traditional self-attention mechanism requires $O(N^2)$ complexity, where N represents the feature length. This corresponds to $O((hwd)^2)$ complexity, where h, w, d

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 171–180, 2022.
https://doi.org/10.1007/978-3-031-21014-3_18

represent the height, width, depth of volume respectively. This is beyond the maximum memory for most current GPUs. This limits the Transformer's application in 3D segmentation for biomedical images. To address this significant burden, some methods have been proposed to reduce the number of tokens for 3D segmentation. For instance, the TransUnet inserted the Transformer block only after the deepest layer of feature extraction in the traditional Unet segmentation structure to model the long-distance dependence [3]. UNETR applied the self-attention after dividing into a sequence of uniform non-overlapping large patches, thus reducing the number of feature lengths [4,16]. However, it is obvious that these segmentation method limits global information modeling only in the higher level feature layer, not otherwise.

In this paper, we propose a new dynamic linear Transformer algorithm for 3D biomedical image segmentation enabling us to apply the Transformer on volumetric data with *linear complexity*. Recently linear computation self-attention could reduce the complexity from $O(N^2)$ to $O(N)$ which dramatically reduced the computation requirement [7,15,18]. Furthermore, the segmentation target is located within a certain physical region rather than the whole frame for many biomedical image segmentation tasks. Thus, almost all previous works used a two-step segmentation strategy to improve the segmentation performance [21]. This guided us to question that *"do we really need to apply self-attention within the whole volume?"* Here, we apply a new dynamic strategy to further limit the computation within the region of interest (ROI). This strategy could further reduce the computation to $(1+\beta)^2\alpha^2$ where α is the relative ratio of ROI size over the whole frame and β is the extension ratio. Besides, by combining the global information the hierarchy decoder structure could simultaneously provide one pixel-level uncertainty map for segmentation.

Clinical Applications: The pancreas is one important abdominal organ that plays an essential role in producing enzymes for digestion and producing insulin and glucagon to adjust blood sugar levels. The segmentation of the pancreas from radiology scans is one crucial work for pancreatic disease diagnosis and treatment [14]. However, this work is very time-consuming and labor-intensive. There are many challenges that limit designing automatic segmentation algorithms for the pancreas. First, the pancreas is a small organ with a sparse and complex shape compared to other abdominal organs. Second, pancreas has one elongating structure which can be hardly defined and long-range dependence is extremely significant for this organ. Third, the contrast of the pancreas in computed tomography (CT) scans is relatively low. Here, we validate our method on two challenging pancreas segmentation datasets to prove the effectiveness of the proposed method.

2 Methods

2.1 Transformer with Linear Complexity

The self-attention mechanism achieved remarkable performance in many computer vision and natural language processing tasks. However, the quadratic complexity prohibits its application in volumetric segmentation tasks. Recently some

methods have been proposed to reduce the computation of self-attention to linear complexity [7,15,18].

Given the individual feature vectors $X \in \mathbb{R}^{n\times d}$, we have the values $V \in \mathbb{R}^{n\times d}$, the queries $Q \in \mathbb{R}^{n\times d}$, the keys $K \in \mathbb{R}^{n\times d}$ after linear projections, where n represents the feature length and d represents the feature dimension. The traditional self-attention mechanism can be expressed in a more general way as follow:

$$V_i^{'} = \sum_{j=1}^{n} sim(Q_i, K_j)V_j \tag{1}$$

where the similarity function is defined as $sim(q,k) = softmax(qk^T/\sqrt{d})$ and i, j are the index for features. In the linear Transformer we want to replace the similarity function such that the similarity function can be divided into two separate parts using the normalized feature quantification function $sim(q,k) = \phi(q)\rho(k)^T$. This replacement could allow us get:

$$V_i^{'} = \sum_{j=1}^{n}(\phi(Q_i)\rho(K_j)^T)V_j = \phi(Q_i)(\sum_{j=1}^{n}\rho(K_j)^T V_j). \tag{2}$$

In this paper, we employ the similarity definition of the Efficient Transformer [15], $\phi(Q_i), \rho(K_j)$ denote applying the softmax function along each row or column of Q, K. This definition allows us to keep the important property of original self-attention that $\sum_{j=1}^{n} sim(Q_i, K_j) = 1$ and reduce the complexity from quadratic to linear.

2.2 Dynamic Transformer

Most Transformer-based image segmentation models are using all the tokens extracted from the raw image. However, we have one question that do we really need all the tokens, and are all the tokens equally meaningful for segmentation? The answer is clearly "no" as for most biomedical image segmentation tasks, the target region usually is located with a certain region due to the physical world limitation. For example, we shall not find tumors located in the kidney if we are looking for liver tumors. Thus the tokens extracted from other regions will be one waste for computation and even further introduce the false positive. In the traditional convolution neural network (CNN), two-stage segmentation methods are widely used to address this problem. Recently, a dynamic Transformer concept is proposed to automatically extract the useful tokens for each input image and reduce the computation for image classification [13]. This inspires us to introduce the dynamic concept from classification into segmentation. We further limit the dense tokens extraction within the ROI regions while in the rest of the image we only take sparse tokens. The ROI region will be decided by the output mask from the previous segmentation layer, which will auto make the whole attention process while keeping the computation linear.

The implementation algorithm of this dynamic sampling is shown in Fig. 1(b). According to the ROI region size, we determine two sampling rates: one high

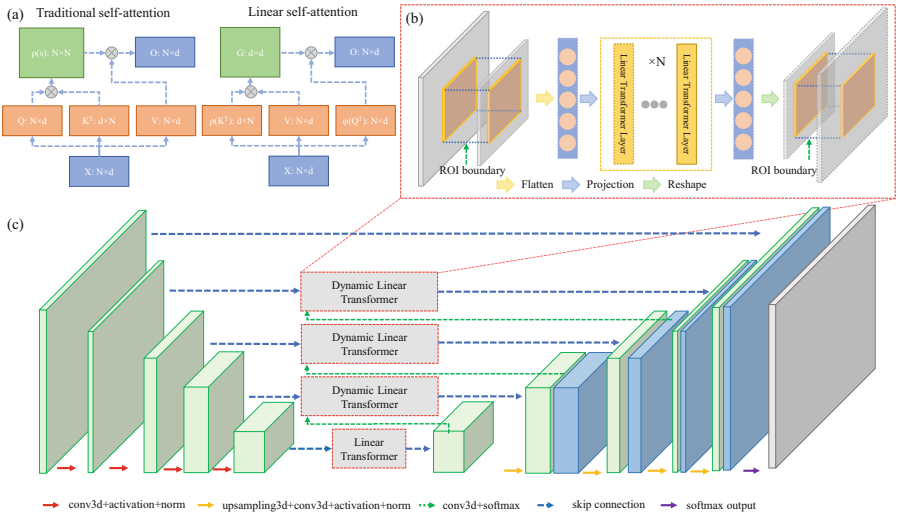


Fig. 1. (a) Comparison of traditional self-attention mechanism v.s. linear self-attention mechanism. (b) The dynamic Transformer block using linear self-attention. (c) The encoder-decoder style segmentation engine.

sampling rate for the ROI region and one low sampling rate for the rest of the region dynamically. For instance, given the ROI size x_0 in the width direction with the whole size of w, we further set two hyper-parameter the relative ROI ratio (α) and extension ratio (β). Then the high sampling rate in the ROI region will be $\frac{\alpha w}{x_0}$, while the sampling rate for the rest of the region will be $\frac{\beta \alpha w}{w - x_0}$. Thus, in this direction the size of the image will be reshaped into $(1 + \beta)\alpha w$ after this ROI resizing process. In this way, we reduce the token numbers for the self-attention calculation to $(1 + \beta)^2 \alpha^2$ of original token numbers. Here we only apply the dynamic ROI resizing in height and widths direction while keeping the original size in the depth direction due to relatively small size in the depth direction. The extracted token after ROI resizing is served into the linear Transformer block to model the long-distance dependence. Next, the image is resized back to its original size according to the spatial position. In this work, we set the $\alpha = 0.5$ according to the pancreas size and $\beta = 0.2$, thus we could reduce the overall computation by 51%. Here these two parameters could be adjusted according to the potential segmentation target size, thus the most of computation is limited within the region of interest rather than the whole image.

2.3 Network Structure

After introducing the linear self-attention and dynamic block, we build one the encoder-decoder style segmentation engineer enhanced with self-attention shown in Fig. 1(c). As shown in the literature [11], the encoded feature map is noisy which limits the performance of self-attention. Thus, the segmented layer also

serves as a spatial attention before the ROI resampling. Beside, similar image embedding and embedding are used in the input and output layer respectively in order to reduce the image size [2]. Other setups are similar to the attention UNet [11]. During the encoder process, we extract the higher-level representative features. In the decoder process, the extracted features are used to generate one segmentation mask at each hierarchy level. As we could expect, the deeper level segmentation loses more detailed information while maintaining one relatively accurate positional information. This positional information can be used in the higher level skip connection to perform the dynamic Transformer block to reduce the computation. In other words, the importance of lower level features will be decided by the deeper level segmentation masks. Thus, in the final output layer, we could combine the accurate positional information coming from the deep feature level and the low-level features to provide one accurate boundary.

3 Results

3.1 Datasets and Training Details

The first pancreas dataset used in the experiments is the NIH pancreas dataset which contains 82 abdominal contrast-enhanced 3D CT scans [14]. These CT scans have resolutions of 512×512 pixels with varying pixel size and slice thickness between 1.5–2.5 mm. The second dataset used in the experiments is the Medical Segmentation Decathlon which contains 282 CT scans with annotation [1]. The training and test dataset is split into 7:1 randomly. Traditional data augmentations including random crop, random rotate, random zoom are employed during the training. We clipped the CT value between -100 and 250 Hu (Hounsfield Unit). Then we normalized the image by extracting the mean value and dividing it by the standard deviation according to the foreground intensity distribution of each pancreas dataset. Note here we directly serve the $512 \times 512 \times 32(Depth)$ volumetric images into the network without slicing into patch.

We implement the proposed network structure using PyTorch with 2 NVIDIA RTX A6000 GPUs. The optimizer is set as the AdamW which is widely used to train the Transformer. The learning rate is set as 0.0001 with one delay rate of 0.5 after 100 epochs and the batch size is set as 8. The training losses are set as the combination of *Dice loss* and *Cross-Entropy loss*, given the pixel position j and the predicted possibility value q_j, the label value p_j:

$$loss = (1 - \frac{\sum_j p_j q_j}{\sum_j p_j + q_j}) + \frac{1}{n}\sum_{j=1}^{n}(p_j log(q_j) + (1 - p_j) log(1 - q_j)). \quad (3)$$

Deep supervision is applied for each decoder layer to guide the training. Thus, the total loss will be the summation of the loss value at each hierarchy stage i,

$$total_loss = \sum_i loss_i. \quad (4)$$

3.2 Segmentation Performance

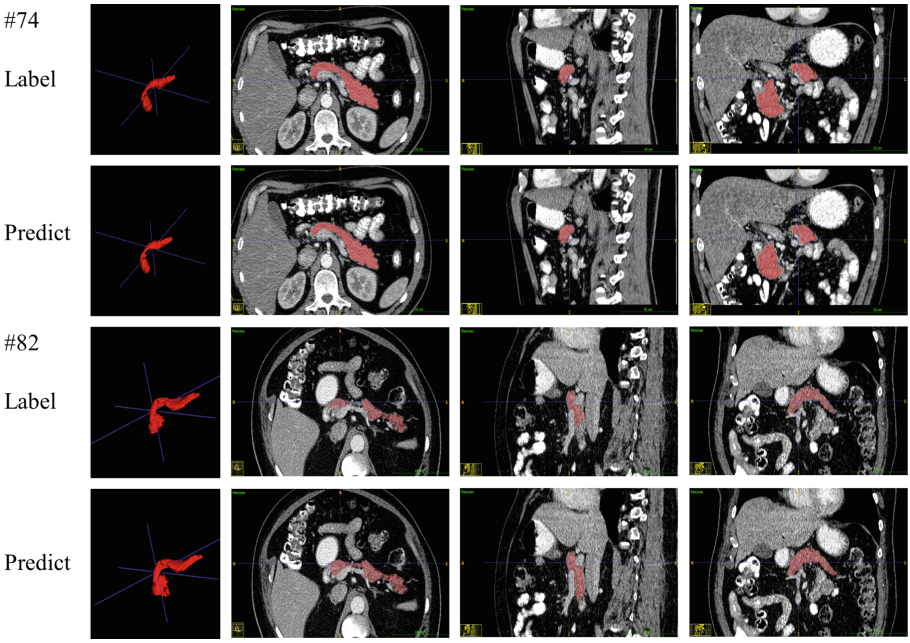


Fig. 2. Some pancreas segmentation examples visualization.

Firstly, we trained our network from scratch on NIH pancreas data. The network converged after about 600 epochs of training. Table 1 showed the performance of our model and the comparison with other methods. Note here all the selected models are trained from scratch without external datasets. One most straightforward comparison is the performance difference between the designed architecture and the attention UNet architecture [11]. Except for the introduction of the linear self-attention mechanism, there is no difference in structure level. We could clearly observe performance improvement by introducing long-distance modeling, which proves the significance of this structure. Figure 2 illustrates two independent segmentation cases in the NIH dataset. The #74 cases achieved one *Dice coefficient* of 0.90 which is higher than the average. The #82 cases achieved one *Dice coefficient* of 0.80 which is lower than the average. We could observe that the proposed model provides one accurate segmentation boundary even compared with the radiologist's annotation, having consistently higher rates with smaller deviations. Particularly, Yu et al. [20] employed a recurrent saliency transformation network to incorporate multi-stage visual cues with coarse segmentation and fine segmentation two stages. Our model surpassed this performance without specific two-stage training. Although fewer researches provide one evaluation metric based on the surface distance, we could observe that our

model achieved one mean surface distance (MSD) of 0.984mm, which indicates the effectiveness of our model.

Table 1. The segmentation performance on the NIH dataset, we employed the Dice coefficient (Dice), Precision, Recall and Mean Surface Distance (MSD) as the evaluation metrics

Methods	Dice	Precision	Recall	MSD (mm)
3D UNet [11]	0.815 ± 0.068	0.815 ± 0.105	0.826 ± 0.062	2.576 ± 1.180
Attention U-Net [11]	0.821 ± 0.057	0.815 ± 0.093	0.835 ± 0.057	2.333 ± 0.856
MDS-Net [8]	0.835 ± 0.062	$\mathbf{0.845} \pm 0.069$	0.837 ± 0.104	–
Liu et al. [9]	0.841 ± 0.049	0.836 ± 0.059	0.853 ± 0.082	–
Yu et al. [20]	0.845 ± 0.050	–	–	–
Ours	$\mathbf{0.855} \pm 0.037$	0.840 ± 0.083	$\mathbf{0.882} \pm 0.040$	$\mathbf{0.984} \pm 0.399$

Secondly, we evaluate our model's performance on the Medical Segmentation Decathlon dataset. We achieved one 0.833 dice coefficient with 0.831 precision and 0.862 recall. This performance further proved that the model could achieve one relatively accurate segmentation on the challenging pancreas segmentation datasets.

3.3 Uncertainty Quantification

Taking advantage of the long-distance modeling ability, the decoded mask at every stage could provide sufficient global information. This allows us to acquire one *uncertainty distribution* over the predictions [5]. Given the predict mask $predict_i$ in different hierarchy stages i and final binary output segmentation map *out*, the definition of the pixel-wise uncertainty is given as follow:

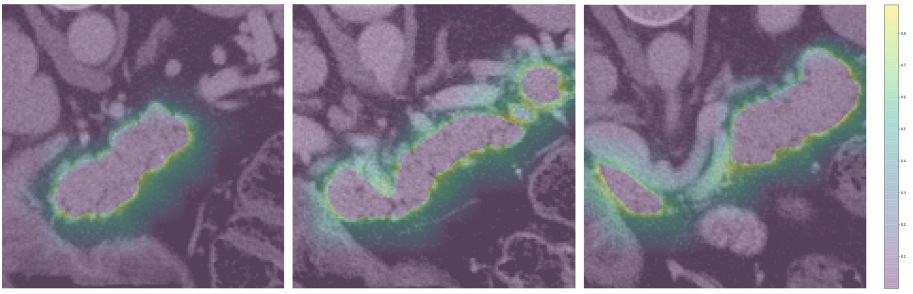


Fig. 3. Uncertainty map visualization, image is cropped according to the pancreas position.

$$Uncertainty = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(predict_i - out)^2}. \quad (5)$$

We anticipate that the high uncertainty value should locate around the boundary. Figure 3 clearly demonstrates that the highest uncertainty values are located around the pancreas boundaries. Furthermore, we calculate the uncertainty value in the true-positive (TP) of 0.040 ± 0.009, true negative (TN) of 0.002 ± 0.000, false negative (FN) of 0.122 ± 0.019, false positive (FP) of 0.563 ± 0.036. We could clearly observe that the mispredicted region has one relatively higher uncertainty value compared with the corrected predicted region with one p value of 5.3×10^{-9}. This agrees with our anticipation that higher uncertainty region is more likely to be mispredicted or, not in agreement with human annotation.

4 Discussion and Concluding Remarks

Although this network structure has achieved promising performance, there are still several improvements that are needed in the future. Particularly on the NIH pancreas dataset, Salanitri et al. [12] achieved current state of the art with one Dice coefficient of 0.880 using hierarchical feature learning. It is well known that the training for self-attention-based structure requires a larger training dataset compared with the traditional CNN network [19], which limits the model's performance on the small dataset like the pancreas segmentation datasets. We could clearly observe that there is one performance gap between the training (Average Dice coefficient: 0.910) and test dataset (Average Dice coefficient: 0.855). A large level pertaining on medical image or external dataset would be helpful to minimize this gap. Some efficient training mechanism designed for Transformer to minimize the data amount requirement is also worth trying in the future [10].

The introduction of global information is essential for 3D biomedical image segmentation. In this study by taking advantage of linear complexity self-attention, we introduced the long-distance modeling into the skip connection between encoder and decoder structure. We showed the feasibility of using linear self-attention for volumetric biomedical image segmentation and proposed one novel dynamic method to further reduce the computation burden. The promising result showed that the self-attention with linear complexity is worth more investigation in the near future for the biomedical image segmentation. In the future we will validate this model on different volumetric biomedical image segmentation including more organs and imaging modalities.

Acknowledgement. This project is funded by the NIH grants: R01-CA246704 and R01-CA240639.

References

1. Antonelli, M., et al.: The medical segmentation decathlon. arXiv preprint arXiv:2106.05735 (2021)
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
3. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
5. Hong, S., et al.: Hypernet-ensemble learning of segmentation probability for medical image segmentation with ambiguous labels. arXiv preprint arXiv:2112.06693 (2021)
6. Huang, X., Deng, Z., Li, D., Yuan, X.: Missformer: an effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162 (2021)
7. Kitaev, N., Kaiser, Ł., Levskaya, A.: Reformer: the efficient transformer. arXiv preprint arXiv:2001.04451 (2020)
8. Li, H., Li, J., Lin, X., Qian, X.: A model-driven stack-based fully convolutional network for pancreas segmentation. In: 2020 5th International Conference on Communication, Image and Signal Processing (CCISP), pp. 288–293. IEEE (2020)
9. Liu, S., et al.: Automatic pancreas segmentation via coarse location and ensemble learning. IEEE Access **8**, 2906–2914 (2019)
10. Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., De Nadai, M.: Efficient training of visual transformers with small-size datasets. arXiv preprint arXiv:2106.03746 (2021)
11. Oktay, O., et al.: Attention U-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
12. Proietto Salanitri, F., Bellitto, G., Irmakci, I., Palazzo, S., Bagci, U., Spampinato, C.: Hierarchical 3D feature learning for pancreas segmentation. In: Lian, C., Cao, X., Rekik, I., Xu, X., Yan, P. (eds.) MLMI 2021. LNCS, vol. 12966, pp. 238–247. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87589-3_25
13. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: efficient vision transformers with dynamic token sparsification. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
14. Roth, H.R., et al.: DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 556–564. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_68
15. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: attention with linear complexities. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3531–3539 (2021)
16. Tang, Y., et al.: Self-supervised pre-training of swin transformers for 3D medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20730–20740 (2022)
17. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: self-attention with linear complexity. arXiv preprint arXiv:2006.04768 (2020)

19. Xu, P., et al.: Optimizing deeper transformers on small datasets. arXiv preprint arXiv:2012.15355 (2020)
20. Yu, Q., Xie, L., Wang, Y., Zhou, Y., Fishman, E.K., Yuille, A.L.: Recurrent saliency transformation network: incorporating multi-stage visual cues for small organ segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8280–8289 (2018)
21. Zhao, N., Tong, N., Ruan, D., Sheng, K.: Fully automated pancreas segmentation with two-stage 3D convolutional neural networks. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 201–209. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_23



Automatic Grading of Emphysema by Combining 3D Lung Tissue Appearance and Deformation Map Using a Two-Stream Fully Convolutional Neural Network

Mohammadreza Negahdar(✉)

Genentech, South San Francisco, CA, USA
negahdar.reza@gene.com

Abstract. We propose and develop an end-to-end learning pipeline incorporating lung tissue appearance and deformation pattern during two phase respiratory CT chest imaging. We formulate the task as a voxel level classification and quantification problem for the emphysematous areas of lung parenchyma on CT imaging and develop a two-stream fully convolutional neural network which fuses together lung tissue pattern and deformation map in a unified framework. The Jacobian matrix, computed between inspiratory and expiratory images, is used as a surrogate measurement of lung deformation. The DenseVNet architecture is adopted for memory efficiency and accuracy on smaller structure segmentation.

The proposed model is trained from inspiratory and expiratory CT images, where the Jacobian matrix calculates the volumetric deformation map of the lung during respiration from inspiration to expiration: the former is used for tissue appearance of the lung while both together provide motion characteristics. Our ablation study shows that incorporating tissue appearance and lung motion accurately classifies all patients with different types and severity of emphysema against the non-emphysema cases. In addition, the proposed method outperforms previous state-of-the-art methods in grading and severity calculation of emphysema, which enables automatic image-based emphysema grading for the first time. To the best of our knowledge, our work is the first proposed framework to incorporate both tissue appearance and deformation map in a deep learning platform for disease diagnosis in lung using CT imaging, which can be intuitively extended to other moving organs and diseases.

Keywords: Lung tissue appearance · Deformation map · Emphysema · Two-stream FCNN

1 Introduction and Related Work

Emphysema is a progressive disease due to permanent enlargement of air spaces distal to terminal bronchioles accompanied by destruction of alveolar walls. Volumetric thoracic computed tomography (CT) imaging allows the study of these morphological changes, and hence the diagnosis and quantification of emphysema [1, 2]. Assessing

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 181–190, 2022.
https://doi.org/10.1007/978-3-031-21014-3_19

the volume and distribution pattern of the emphysematous tissue is needed to evaluate disease severity and progression, and to predict future risks such as lung cancer [3–6] as well as potential vulnerability and complications to certain epidemic and/or pandemic diseases.

Since restrictive and obstructive lung diseases are manifested as textural alterations in the lung parenchyma, convolutional neural networks (CNNs) and deep learning models show tremendous potential for learning features to classify different lung tissue pathology [7–11]. However, most studies have abated the problem of emphysema detection by measuring low-attenuation areas (LAA) of the lung [9, 12, 13] while accurate quantification and severity assessment are an essential part of treatment planning and longitudinal tracking of lung diseases.

Novel therapeutic approaches for lung disease require intricate patient selection. This requires expert radiologists to visually and quantitatively evaluate thoracic CT images, along with additional functional acquisitions such as paired inspiratory-expiratory scans [14]. Hence lung tissue can be thoroughly characterized and distinguished based on its motion characteristics [15–17].

In this study, we propose and develop an end-to-end learning pipeline to incorporate lung tissue appearance and deformation pattern during respiration. We aim to delineate emphysematous areas of lung for diagnosis and severity grading. We develop a two-stream fully convolutional neural network (FCNN) which fuses together lung tissue pattern and deformation map in a unified framework to thoroughly characterize lung tissue.

The idea of combining motion and appearance together has been already studied in video classification [18], object tracking [19], and video object segmentation [20, 21]. While we got inspired by these studies, we present the first deep learning framework to combine volumetric lung tissue appearance and deformation map using a two-stream FCNN for lung tissue characterization. We strongly believe that findings of this study will ignite new ideas to grow this concept in the medical imaging community.

We extensively evaluated the performance of the proposed pipeline. We also provided an ablation analysis to investigate the influence of each stream in the proposed framework.

The contributions of the paper are two-fold:

we present a robust and fully automatic end-to-end framework for the diagnosis and severity grading of emphysema using only chest CT images, for the first time.
To the best of our knowledge, our work is the first proposed framework to leverage motion characteristics along with tissue appearance using a two-stream FCNN for disease diagnosis and severity grading in medical imaging.

2 Approach

We aim to accurately segment and quantify emphysematous areas of lung by evaluating the synergy of lung tissue appearance and deformation map, as shown in Fig. 1. In an ablation study, first, we carry out the problem using the inspiratory image for lung tissue appearance only. Then, we use both inspiratory and expiratory images of the lung during

a respiration cycle to address the problem by deformation map information. Ultimately, we fuse the two streams to solve the problem using the joint model.

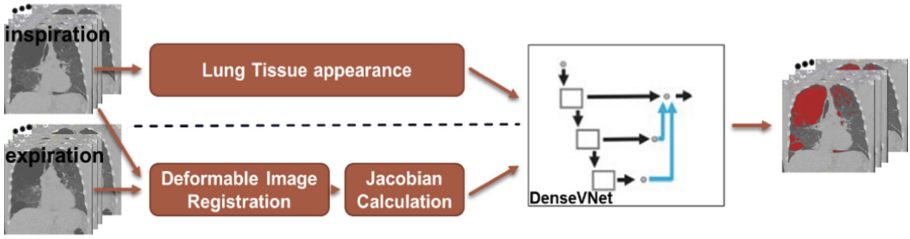


Fig. 1. Pipeline structure of our proposed model. Lung tissue appearance (stream above dot line) and lung motion characteristics (stream below dot line) incorporate in a joint model to characterize lung tissue for emphysema detection

2.1 Architecture

In order to capture high-resolution data of thoracic CT images and encode spatial and temporal features for dense labeling of voxels, computational and memory cost should be considered. Dense networks reuse feature representations multiple times [22], and in conjunction with the multi-scale VNet architecture [23, 24] significantly improves the performance. Dense feature stacks and batch-wise spatial dropout yield deeper networks at higher resolutions, which enables the segmentation of smaller areas and improved memory efficiency [25, 26]. We adopted a Tensorflow implementation of a pre-trained DenseVNet in NiftyNet, where each downsampling stage of VNet architecture is a dense feature stack while upsampling stage comprises bilinear upsampling and skip connections are convolutions. The modified Dice loss is adopted as loss function [25]. We used dilated convolutional layers to improve network's ability to detect both large and small areas [22].

Lung Tissue Appearance Stream
This stream produces a high-spatial resolution of features and high-level encoding of lung tissue pattern through dilated convolutions, which is a distinctive feature to characterize emphysematous tissue against other lung tissue types and normal areas of lung [11]. Spatial stream analyzes individual inspiratory image which is clinically studied by radiologists for LAA detection.

Lung Tissue Motion Stream
Lung diseases change the biomechanical specification [17] and deformation map of the lung during respiration [16], as shown in Fig. 2. The right panel of Fig. 2 shows how lung inflammation (yellow arrows) tinged the color-coded deformation map with blue in left lung and abruptly highlighted the upper lobe in the right lung. We exploited the Elastix toolbox for consecutive affine and deformable image registration (DIR) of lung during respiration [27], computed between inspiratory and expiratory images. Affine

registration addresses patient motion artifact. The determinant of Jacobian of the calculated volumetric deformation map, represents volume changes of lung tissue at each voxel, is used as a surrogate measurement of the lung deformation map.

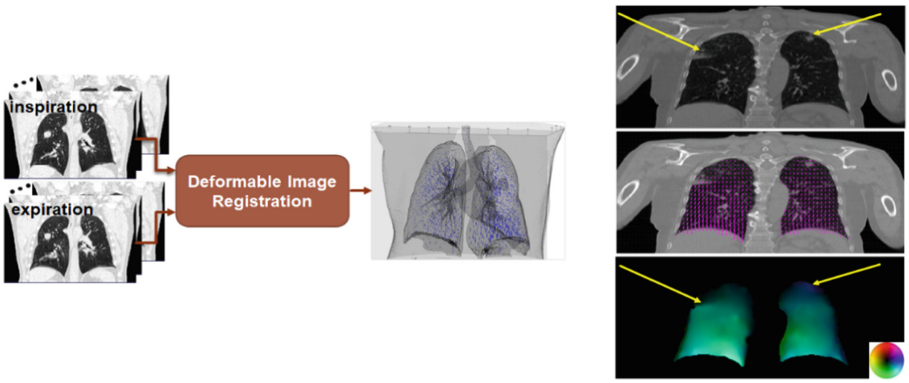


Fig. 2. (Left diagram) Deformation map calculated by DIR method. (Right images): Mid-coronal plane of lung (up) and a vector field (middle) along with a color-coded representation of the calculated deformation map (down), where smaller magnitude of motion is darker while color indicates direction of motion. Yellow arrows point to inflammation area of lung and its different color pattern at color-coded deformation map representation (Color figure online)

There is an opportunity to extend this pipeline to a fully deep learning-based implementation. However, training data scarcity for the lung deformation map led us to adopt DIR method for temporal stream creation.

Fusion Model

The two streams of the network encode spatial and temporal features, which are merged into a lung tissue characterization. We aim to combine the outputs of lung tissue appearance and deformation map streams to leverage a whole that is greater than the sum of its components.

In computer vision community, different streams independently do their task and their outputs fuse together late in the model using either mean fusion [28] or union fusion [29]. We chose to fuse the two streams in early stage, which concatenates the intensity image and Jacobian map into a single channel and feed it to the learning network as a whole. Early fusion can model the low-level interactions between the two streams and eliminate the necessity of training two separate networks.

Considering scarcity of training data, we extensively augmented our training pool by applying random affine transformations and elastic deformation to attain robustness and higher precision on the test dataset.

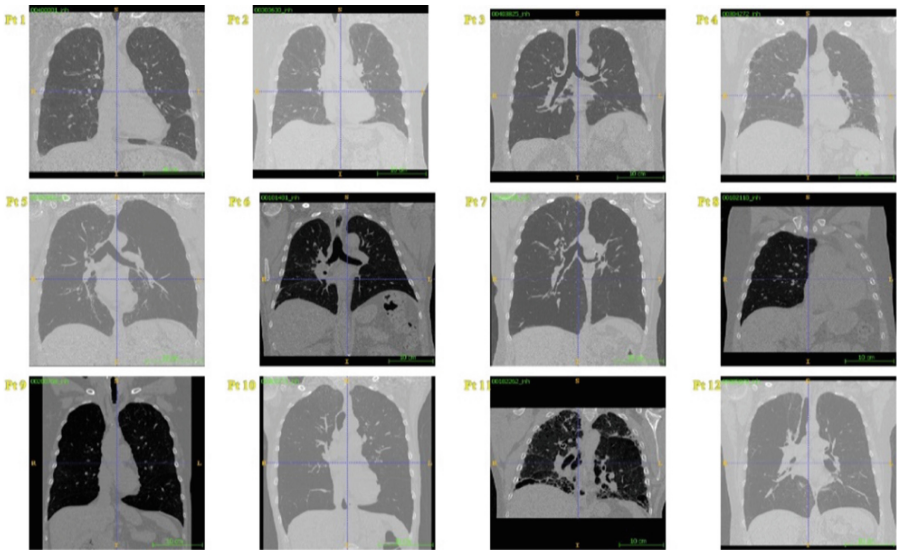


Fig. 3. Twelve patients with different extent and severity of emphysema for performance evaluation of the proposed model

2.2 Evaluation and Clinical Metrics

Emphysema distribution is associated with chronic obstructive pulmonary disease (COPD) [30] and the pulmonary function test (PFT) is the traditional gold standard for COPD diagnosis [31]. Imaging-based diagnosis and severity assessment of emphysema is correlated with COPD prognostic factors such as Tiffeneau-Pinelli index (TI) which calculated as the ratio of forced expiratory volume in one second (FEV1) to total volume of air that the patient can exhale in one breath (FVC). Based on GOLD guidelines, a TI of <0.7 after bronchodilator is considered diagnostic of COPD.

The severity of emphysema can be graded into five stages; each patient is assigned a categorical grade based on the percentage of emphysematous areas in the lung. Patients are graded from 1 to 5 corresponding to 1%–5%, 6%–25%, 26%–50%, 51%–75%, and 76%–100% emphysematous lung volume, respectively [32]. In case of finding no emphysema, patient is graded 0. Although the severity grading of emphysema shows strong relation with COPD GOLD staging, COPD severity is influenced by both chronic bronchitis and emphysema [31, 33], therefore not a perfect correlation.

Quantitative performance of our framework is evaluated against the aforementioned clinical metrics.

Table 1. Comparison of each stream performance against joint model to diagnose and quantify emphysematous areas CTSCANRPT shows primary diagnosis of CT scan by radiologists. Underlined patients (Pts 3, 6, and 8) have no emphysema. Emphysema severity grading is assessed according to Sect. 2.2

	Ground Truth			appearance		motion		joint	
Pt	CTSCANRPT	Emphy%	Grade	Emphy%	Grade	Emphy%	Grade	Emphy%	Grade
1	Centrilobular Emphysema	8.1	2	21.8	2	27.8	3	22.4	2
2	Centrilobular Emphysema	29.1	3	8.6	2	50.9	3	34.1	3
3	Primary Bronchogenic Carcinoma	0.6	0	6	2	4.7	2	0.4	0
4	Centrilobular Emphysema	1.5	1	6.5	2	5.3	2	3.9	1
5	Centrilobular Emphysema	55	4	23.6	2	26.6	3	59.4	4
6	Primary Bronchogenic Carcinoma	0	0	6.4	2	3.1	1	0.7	0
7	Centrilobular Emphysema	28.6	3	11.8	2	41.3	3	34.4	3
8	Primary Bronchogenic Carcinoma	0	0	0	0	2.3	1	0.7	0
9	Centrilobular Emphysema	54.9	4	31.4	3	54.1	4	56.3	4
10	Panlobular Emphysema	53.9	4	10	2	20.5	2	66.4	4
11	Centrilobular Emphysema	4.3	1	0	0	19.4	2	4.8	1
12	Centrilobular Emphysema	45.8	3	21.7	2	37.6	3	47.8	3

3 Results and Discussions

3.1 Dataset

We used the lung tissue research consortium (LTRC) dataset, which provides a dense interstitial lung disease (ILD) labeling of the entire lung [34]. The LTRC dataset consists of patients of different restrictive and obstructive lung diseases. We collected 120 patients who have both inspiratory and expiratory CT images along with labeling on inspiratory CT images. These patients have different extent of all three morphological types of emphysema: centrilobular, paraseptal, and panlobular. We use 70% of patients (84 patients) for training our proposed model, 20% of patients (20 patients) for model validation, and 10% of patients (12 patients) are reserved for performance evaluation, as shown in Fig. 3.

3.2 Implementation

Our model is trained by minimizing the modified dice loss with a learning rate of 10^{-5}. We trained the model for 500 epochs using Keras with a TensorFlow backend and the learning rate is gradually decreased. We trained our model using four NVIDIA Tesla P100 GPU with 16 GB of memory.

3.3 Ablation Study

To evaluate the influence of each stream and performance of the joint model, in an ablation study, we trained the proposed model with each stream separately, and ultimately jointly. Table 1 demonstrates the results of the ablation study against ground truth. The joint model correctly diagnoses all emphysema against non-emphysema cases: Patients 3, 6, and 8 (Grade 0).

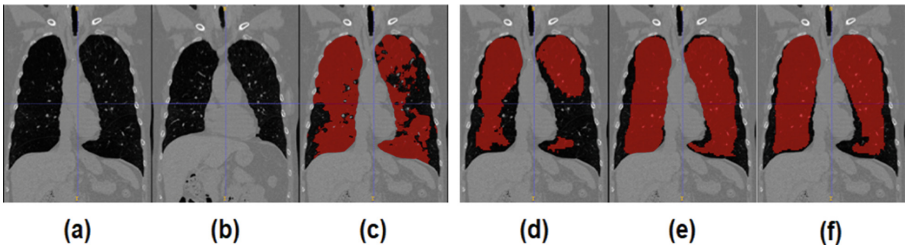


Fig. 4. Patient 9 with severe lung emphysema: (a) inspiratory, (b) and expiratory image along with (c) delineated emphysematous area by radiologist. Calculated emphysematous area by (d) appearance, (e) motion, and (f) joint model

Results show that appearance model generally underdiagnose emphysematous areas which might be because of mild or early stages of emphysema where there is not enough spatial representation. In addition, we have not had enough patients with some specific types of emphysema such as paraseptal emphysema, which can be addressed by creating a bigger pool of training data.

Results of motion stream show more volatility, which might be caused by the lack of a repeatable breathing protocol. As shown in last column of Table 2, patients have different depth of inhalation, which potentially cause inaccuracies in DIR output, consequently motion stream.

The joint model successfully assessed severity grade of emphysema for all twelve patients, which shows the distinct advantage of learning from both signals in a unified framework: a true synergy, with substantially stronger results than what we can obtain from either one alone. Figure 4 shows input and output of aforementioned three trained models of ablation study for patient nine.

Based on the PFT results presented in Table 2, diagnosis and severity grading of emphysema are correlated, not linearly indicative of TI and COPD GOLD staging, though. It might be because TI and GOLD are directly indicative of COPD severity which influenced by both chronic bronchitis and emphysema. Thus, our proposed model for emphysema diagnostics and severity grading could contribute to complementary phenotyping as part of CT imaging biomarker for patients with COPD in combination with PFT to assess bronchitis severity.

4 Conclusions

The majority of visual, semiquantitative, and quantitative emphysema assessment techniques have heavily relied on radiologists, and are subject to increased interobserver and interpatient variability. Therefore, offering a robust and fully automated framework for image-based diagnosis and quantification of emphysema severity has a big value and clinical impact. To the best of our knowledge, this is the first proposed framework to leverage motion characteristics along with lung tissue appearance using a two-stream FCNN for disease diagnosis and severity grading in medical imaging.

Due to the gradual onset of emphysema, one can develop emphysema for several years without noticing any signs or symptoms. Therefore, our model might address the

Table 2. PFT result of all twelve patients picked for performance evaluation of the proposed method last three columns present total lung volume (TLC), residual volume (RV) and inflation rate during respiration as a representation of inhalation depth

Pt	age	postFEV1predict	postFVCpredict	TI index	postGOLD	TLC	RV	(TLC-RV)/RV
1	46	23	43	0.43	4			
2	56	27	102	0.27	4	7.88	4.91	60
3	65	51	93	0.55	2	6.56	3.53	85
4	60	85	107	0.79	0	4.1	1.35	203
5	48	14	68	0.21	4	7.57	5.41	39
6	64	101	122	0.83	0	6.09	2.14	184
7	60	24	111	0.21	4	8.43	5.12	64
8	61	44	63	0.7	0	4.88	2.89	68
9	55	21	62	0.34	4	6.98	4.47	56
10	52	20	65	0.25	4			
11	60	84	108	0.78	0	5.34	1.7	214
12	52	15	57	0.21	4			

challenging scenario where the disease at early stages has no symptoms or appearance manifestation, but potentially affects motion characteristics of the lung tissue.

Since our approach is not specific to a particular network, organ, or imaging modality, it could be a general framework for abnormality detection in any moving organ in the body, providing scope for future expansion of the concept.

References

1. Fernandes, L., Gulati, N., Mesquita, A.M., Sardesai, M., Fernandes, Y.: Quantification of emphysema in chronic obstructive pulmonary disease by volumetric computed tomography of lung. Indian J. Chest Dis. Allied Sci. **57**, 155–160 (2015)
2. Omori, H., Fujimoto, K., Katoh, T.: Computed-tomography findings of emphysema: correlation with spirometric values. Curr. Opin. Pulm. Med. **14**, 110–114 (2008)
3. Wille, M.M.W., et al.: Visual assessment of early emphysema and interstitial abnormalities on CT is useful in lung cancer risk analysis. Eur. Radiol. **26**(2), 487–494 (2015). https://doi.org/10.1007/s00330-015-3826-9
4. Lynch, D.A., Austin, J.H., Hogg, J.C., Grenier, P.A., Kauczor, H.U., et al.: CT-definable subtypes of chronic obstructive pulmonary disease: a statement of the fleischner society. Radiology **277**, 192–205 (2015)
5. Hohberger, L.A., Schroeder, D.R., Bartholmai, B.J., Yang, P., Wendt, C.H., et al.: Correlation of regional emphysema and lung cancer: a lung tissue research consortium-based study. J. Thorac. Oncol. **9**, 639–645 (2014)
6. Haruna, A., Muro, S., Nakano, Y., Ohara, T., Hoshino, Y., et al.: CT scan findings of emphysema predict mortality in COPD. Chest **138**, 635–640 (2010)
7. Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., Mougiakakou, S.: Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans. Med. Imaging **35**, 1207–1216 (2016)
8. Gao, M., Bagci, U., Lu, L., Wu, A., Buty, M., et al.: Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. Comput. Methods Biomech. Biomed. Eng. Imaging Visual. **6**, 1–6 (2018)
9. Ørting, S. N., Petersen, J., Thomsen, L. H., Wille, M. M. W., and Bruijne, M. d.: Detecting emphysema with multiple instance learning. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 510–513. (2018)

10. Christodoulidis, S., Anthimopoulos, M., Ebner, L., Christe, A., Mougiakakou, S.: Multisource transfer learning with convolutional neural networks for lung pattern analysis. IEEE J. Biomed. Health Inform. **21**, 76–84 (2017)
11. Negahdar, M., Beymer, D.: Lung tissue characterization for emphysema differential diagnosis using deep convolutional neural networks. In: SPIE Medical Imaging, vol. 10950, pp. 109503R. San Diego, CA (2019)
12. Bermejo-Peláez, D., Estepar, R.S.J., Ledesma-Carbayo, M.J.: Emphysema classification using a multi-view convolutional network. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 519–522 (2018)
13. Fischer, A.M., Varga-Szemes, A., van Assen, M., Griffith, L.P., Sahbaee, P. et al.: Comparison of artificial intelligence–based fully automatic chest CT emphysema quantification to pulmonary function testing. Am. J. Roentgenol. **214**(5), 1065–1071 (2020)
14. Kauczor, H.-U., Wielpütz, M.O., Jobst, B.J., Weinheimer, O., Gompelmann, D., et al.: Computed tomography imaging for novel therapies of chronic obstructive pulmonary disease. J. Thorac. Imaging **34**, 202–213 (2019)
15. Negahdar, M., Amini, A.A.: Regional lung strains via a volumetric mass conserving optical flow model. In: IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro, pp. 1475–1478. Barcelona, Spain (2012)
16. Negahdar, M., Dunlap, N., Zacarias, A., Civelek, A.C., Woo, S.Y. et al.: Comparison of indices of regional lung function from 4-D X-ray CT: Jacobian vs. strain of deformation. In: IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro, pp. 648–651. San Francisco, CA, USA (2013)
17. Negahdar, M., Fasola, C.E., Yu, A.S., von Eyben, R., Yamamoto, T., et al.: Noninvasive pulmonary nodule elastometry by CT and deformable image registration. Radiother. Oncol. **115**, 35–40 (2015)
18. Joe Yue-Hei, N., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R. et al.: Beyond short snippets: deep networks for video classification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4694–4702 (2015)
19. Wang, L., Ouyang, W., Wang, X., Lu, H.: Visual tracking with fully convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3119–3127 (2015)
20. Jain, S.D., Xiong, B., Grauman, K.: FusionSeg: learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117–2126 (2017)
21. Li, S., Seybold, B., Vorobyov, A., Lei, X., Kuo, C.C.J.: Unsupervised video object segmentation with motion-based bilateral networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., et al. (eds.) Computer Vision – ECCV 2018, pp. 215-231. Cham (2018)
22. Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., et al.: Towards image-guided pancreas and biliary endoscopy: automatic multi-organ segmentation on abdominal CT with dense dilated networks. In: Medical Image Computing and Computer Assisted Intervention − MICCAI 2017, pp. 728-736. Cham (2017)
23. Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on, pp. 565–571 (2016)
24. Negahdar, M., Beymer, D., Syeda-Mahmood, T.: Automated volumetric lung segmentation of thoracic CT images using fully convolutional neural network. In: SPIE Medical Imaging, vol. 10575 (2018)
25. Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., et al.: Automatic multi-organ segmentation on abdominal CT with dense V-networks. IEEE Trans. Med. Imaging **37**, 1822–1834 (2018)

26. Ghavami, N., Hu, Y., Gibson, E., Bonmati, E., Emberton, M., et al.: Automatic segmentation of prostate MRI using convolutional neural networks: investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration. Med. Image Anal. **58**, 101558 (2019)
27. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.: elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**, 196–205 (2010)
28. Kalogeiton, V., Weinzaepfel, P., Ferrari, V., Schmid, C.: Action tubelet detector for spatio-temporal action localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 4415–4423 (2017)
29. Singh, G., Saha, S., Sapienza, M., Torr, P., Cuzzolin, F.: Online real-time multiple spatiotemporal action localisation and prediction. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3657–3666 (2017)
30. Mohamed Hoesein, F.A., van Rikxoort, E., van Ginneken, B., de Jong, P.A., Prokop, M., et al.: Computed tomography-quantified emphysema distribution is associated with lung function decline. Eur. Respir. J. **40**, 844–850 (2012)
31. Bailey, K.L.: The importance of the assessment of pulmonary function in COPD. Med. Clin. North Am. **96**, 745–752 (2012)
32. Wille, M.M.W., Thomsen, L.H., Dirksen, A., Petersen, J., Pedersen, J.H., Shaker, S.B.: Emphysema progression is visually detectable in low-dose CT in continuous but not in former smokers. Eur. Radiol. **24**(11), 2692–2699 (2014). https://doi.org/10.1007/s00330-014-3294-7
33. Heussel, C.P., Herth, F.J.F., Kappes, J., Hantusch, R., Hartlieb, S., et al.: Fully automatic quantitative assessment of emphysema in computed tomography: comparison with pulmonary function testing and normal values. Eur. Radiol. **19**, 2391–2402 (2009)
34. Bartholmai, B., Karwoski, R., Zavaletta, V., Robb, R., Holmes, D.: The lung tissue research consortium: an extensive open database containing histological, clinical, and radiological data to study chronic lung disease. In: The Insight Journal—2006 MICCAI Open Science Workshop (2006)



A Novel Two-Stage Multi-view Low-Rank Sparse Subspace Clustering Approach to Explore the Relationship Between Brain Function and Structure

Shu Zhang[1(✉)], Yanqing Kang[1], Sigang Yu[1], Jinru Wu[1], Enze Shi[1], Ruoyang Wang[1], Zhibin He[2], Lei Du[2], and Tuo Zhang[2]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
shu.zhang@nwpu.edu.cn

[2] School of Automation, Northwestern Polytechnical University, Xi'an, China

Abstract. Understanding the relationship between brain function and structure is vital important in the field of brain image analysis. It elucidates the working mechanism of the brain, which will contribute to better understand the brain and simulate the brain-like system. Extensive efforts have been made on this topic, but still far from the satisfactory. The major difficulties are at least two aspects. One is the huge individual difference among the subjects, which makes it hard to obtain stable results at groupwise level, e.g., noise signals can significantly affect the exploring process. The other one is the huge difference between functional and structural features of the brain, both in their pattern and size, which are very different. To alleviate the above problems, in this paper, we propose a two-stage multi-view low-rank sparse subspace clustering (Two-stage MLRSSC) method to jointly study the relationship between brain function and structure and identify the common regions of brain function and structure. The major innovation of proposed Two-stage MLRSSC is that comparable features of brain function and structure can be effectively extracted from low-rank sparse representation, and results are further improved the stability by two-stage strategy. Finally, groupwise-based stable functional and structural common regions are identified for better understanding the relationship. Experimental results shed new ways to explore the brain function and structure, new insights are observed and discussed.

Keywords: Brain function · Brain structure · Multi-View · Low-Rank Sparse Subspace Clustering

1 Introduction

Understanding the relationship between brain function and structure is the key issue to study the working mechanism of brain. To this end, extensive research has been proposed to study functional connectivity (FC), structural connectivity (SC), and their relationships [1, 2]. In the early stage, brain function and structure were usually investigated solely, since brain function is complex with huge dynamic changes, while brain

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 191–200, 2022.
https://doi.org/10.1007/978-3-031-21014-3_20

structure is unique and relatively immutable [3, 4]. Later on, many methods have been proposed to correlate different modality profiles of the brain, such as PCA and CCA [5]. In recent years, multi-modality analysis is becoming a very hot topic in the medical image field. With the help of those powerful algorithms, brain function and structure could be effectively associated and analyzed, e.g., susceptible-infected-susceptible (SIS) algorithm was used to analyze the relationship between function and structure among brain regions [6], and Liang et al. analyzed their relationship through the topological mapping of structure-function connection [7].

Despite the great success above-mentioned approaches already achieved, they are still far from the satisfactory for the researchers to study the relationship between brain function and structure. On one hand, the structural and functional features of the brain are both complex and usually hidden in the form of higher dimensions, and the extracting process is easy to be affected by the noises or individual differences [8]. On the other hand, since functional and structural features are very different in many perspective, certain analysis should be arranged first to put them under the same space [9].

To alleviate the above problems, in this paper, we propose a two-stage multi-view low-rank sparse subspace clustering (Two-stage MLRSSC) method to jointly learn a subspace representation by constructing affinity matrix shared by brain functional and structural representations. The major innovation is that proposed method encourages low-rank and sparsity of the feature representation [10] and subspace learning can deal with highly heterogeneous multi-view representation and is very suitable for joint analysis of brain functional and structural features with great differences. Low-rank and sparse representations can simultaneously extract global and local features of brain and obtain effective low-dimensional representations of brain function and structure. In addition, Human brain owns huge differences across the individuals, in order to explore the stable relationship, group analysis is introduced into the traditional MLRSSC method, which further improve learning the stable relationship from individuals.

Experimental results successfully demonstrate the effectiveness of proposed method. By clustering, the brain surface can be adaptively divided into several common regions, such as 10 or 20. The relationship between brain function and structure was identified and localized through the analysis of common regions. Besides, similar phenomenon and conclusions are found by comparing our results with gradient in cortical features [11], which to show how the spectrum of cortical function emerges from structural constraints. And hierarchical characteristics can also be observed through exploring the different scale of the common regions.

2 Method

2.1 Overview of the Proposed Two-Stage MLRSSC

The pipeline of proposed Two-stage MLRSSC method is shown in Fig. 1. For each subject, its cortical surface is initialized to 150 Destrieux regions (DRs) using the Destrieux Atlas [12]. Based on the initialized region, each subject's whole brain fiber connections are utilized to generate the structural feature matrix, and rs-fMRI is utilized to generate the functional feature matrix.

A two-stage MLRSSC is proposed to achieve clustering at the group level instead of the individual level. As shown in Fig. 1(A), in the first stage, we use MLRSSC method on 100 individuals to obtain the individual surface partitioning results; in the second stage, the surface partition template is obtained based on the group clustering approach. Specifically, in the second stage, we first divided 100 individuals into several sets. Then, set-inner surface partition results are generated by combining individuals of the same set. Finally, the group surface partitioning template is generated by combining all the set-inner surface partition results.

The group clustering in Fig. 1(B) describes the process of combining a set of individuals to produce a result. Final result is obtained by a kind of iterative optimization way. For the t_{th} iteration, we have the weight coefficients $\mathbf{A}^t$ corresponding to the feature matrices of the individuals. The weighted feature matrices $\boldsymbol{F}^t$ and $\boldsymbol{S}^t$ are obtained by weighting the feature matrices of individuals with $\mathbf{A}^t$. Then we obtain the clustering result by MLRSSC for the weighted feature matrices. In order to obtain better weight coefficients and clustering result, we design the optimization objective for iterative optimization. The optimization objective is to minimize the average similarity between the clustering result and the individual results.

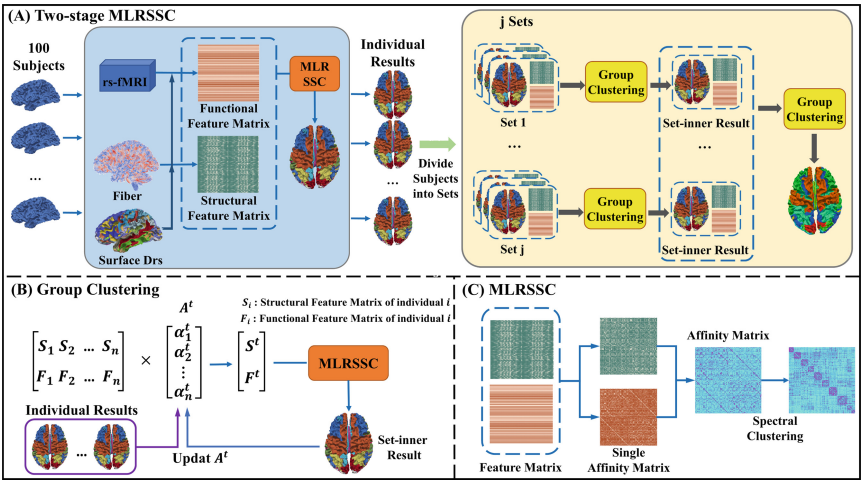


Fig. 1. The pipeline of proposed Two-stage MLRSSC. (A) represents the proposed method process. In the first stage, MLRSSC is used for individual, and in the second stage, Group Clustering is used for groups to get cortical surface partition template. (B) shows method Group Clustering. A^t represent the weight coefficient of the t_{th} iteration. (C) shows method MLRSSC. (Color figure online)

2.2 Data and Preprocess

100 subjects were randomly selected from the Human Connectome Project (HCP) 900 dataset [13]. T1-weighted magnetic resonance imaging (MRI) data were used to reconstruct the brain cortical surface, diffusion-weighted MRI (dMRI) were utilized to reconstruct the fiber bundles from white matter and resting-state functional MRI (rs-fMRI) showed the functional changes of the brain via fMRI time series.

Standard Freesurfer pipeline including tissue-segmentation and white matter surface (inner surface) reconstruction [14, 15] is proposed to preprocess the T1-weighted MRI. For rs-fMRI, we adopted the graycoordinate system [16] as the platform to extract rs-fMRI time sequence for each surface vertex. For dMRI, we followed the method in Van den Heuvel et al. [17] to use deterministic tractography to derive white matter fibers, and reconstructed 5×10^4 fiber tracts for each subject.

To jointly use these three modalities, we aligned them into the same space. A linear image registration method (FLIRT) [18] and a nonlinear one (FNIRT) [19] were cascaded to transform and warp T1-weighted MRI to the dMRI space. The preprocessed rs-fMRI uses graycoordinate system, which is located at the same space as T1-weighted MRI, and the vertex-wise correspondence between them can be directly established.

2.3 Brain Functional and Structural Profiles

To jointly study the brain function and structure, it is important to generate the brain functional and structural features. For the individual, we generated its feature matrices based on 150 brain surface DRs.

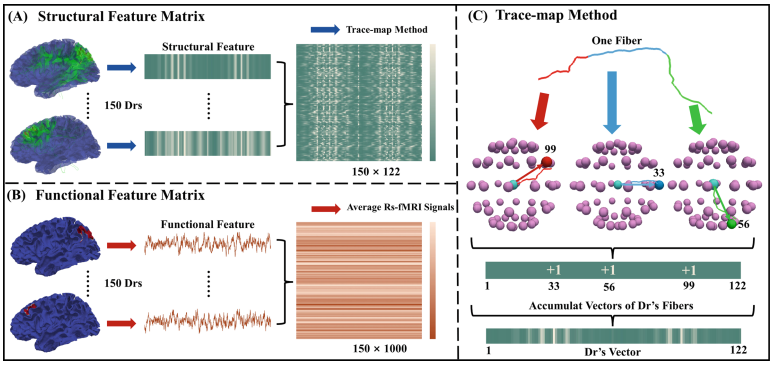


Fig. 2. Brain functional and structural features extraction. (A) and (B) show the process of extracting the subject structural and functional feature matrices, respectively. (C) shows the Trace-map method utilizing fibers as features.

The fiber connections patterns are chosen to present the structural features, as shown in Fig. 2(A). For certain DR, fibers, which are passing through it, are picked up and used to generate its structural feature by Trace-map [20] method whose details are shown in Fig. 2(C). Specifically, for each fiber, three major fiber-directions are calculated and arranged into certain regions (colorful balls). Since the whole sphere is divided into 122 regions without overlap, a vector of 122 elements can be used to represent the shape of a fiber [20]. Thus, the fiber bundles can be represented by accumulating the vectors from each fiber. Rs-fMRI signals are used to represent functional features. For certain DR, the average of all its rs-fMRI signals, which are time sequences with a length of 1000, is used as its functional feature, as shown in Fig. 2(B).

Finally, we get a set of brain structural feature matrices with dimensions of 150 × 122 and functional feature matrices with dimensions of 150 × 1000.

2.4 Two-Stage MLRSSC

For each subject, the structural and functional feature matrices based on 150 brain surface DRs were obtained. The MLRSSC was used to combine structural and functional feature matrices to cluster 150 DRs on the brain surface, so to gather the regions with strong consistency and achieve the partition of brain surface common regions.

In MLRSSC, a low-rank sparse representation matrix is obtained for every view. Then all the representation matrices are combined to obtain the affinity matrix. Finally, the affinity matrix is used as the input for spectral clustering to obtain the clustering result. The affinity matrix $\boldsymbol{W}$ is Eq. (1)

$$
\begin{aligned}
&W = \left|C_{avg}\right| + \left|C_{avg}\right|^T \\
&C_{avg} = \frac{\sum_{v=1}^{n_v} C^{(v)}}{n_v} \\
&\min_{C^{(1)}C^{(2)},\ldots,C^{(n_v)}} \sum_{v=1}^{n_v} \left(\beta_1 \left\|C^{(v)}\right\|_* + \beta_2 \left\|C^{(v)}\right\|_1\right) + \sum_{1\le v,w\le n_v, v\ne w} \lambda^{(v)} \left\|C^{(v)} - C^{(w)}\right\|_F^2 \\
&s.t.\quad X^{(v)} = X^{(v)}C^{(v)}, diag\left(C^{(v)}\right) = 0, v = 1,\ldots,n_v
\end{aligned}
\tag{1}
$$

where $\boldsymbol{C}^{(v)}$ is the representation matrix for view v. n_v is the number of views for data $\boldsymbol{X}$. Parameters β_1, β_2 and $\lambda^{(v)}$ define the trade-off between low-rank, sparsity constraint and the agreement across views, respectively. We set $\beta_1 = 0.3$, $\beta_2 = 0.7$, $\lambda^{(v)} = 0.3$ for all views $v = 1, \ldots, n_v$ in empirically.

Algorithm 1 Two-Stage MLRSSC

Input: $\boldsymbol{X}=\{\boldsymbol{X}_i\}_{i=1}^{n}$, $\boldsymbol{X}_i=\{\boldsymbol{X}_i^{(v)}\}_{v=1}^{n_v}$
Output: Group surface partition template
for $i = 1$ **to** n **do**
Obtain the clustering matrix $\boldsymbol{L}_i$ of each subject $\boldsymbol{X}_i$ with MLRSSC
end for
Calculate the similarity matrix $\boldsymbol{S}_{n*n}$ with the SSIM [21] between clustering matrices$\{\boldsymbol{L}_i\}_{i=1}^{n}$
Divide n subjects into j sets $\{\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots \ldots, \boldsymbol{C}_j\}$ $\boldsymbol{s.t.}$ $\boldsymbol{C}_c = \{\boldsymbol{X}_i\}_{k_c(1)}^{k_c(n_c)}$ with $\boldsymbol{S}_{n*n}$
for $c = 1$ **to** j **do**
Generate a set-inner feature matrix $\boldsymbol{P}_c$ from subjects in set $\boldsymbol{C}_c$ using Particle Swarm Optimization (PSO) Algorithm [22] with function $f(\boldsymbol{P}_c)$, $\boldsymbol{L}_{\boldsymbol{P}_c}$ is the clustering matrix of $\boldsymbol{P}_c$ with MLRSSC

$$\boldsymbol{P}_c = \sum_{z=1}^{n_c} \boldsymbol{\alpha}_z \boldsymbol{X}_{k_c(z)} \;\; s.t. \;\; \sum_{z=1}^{n_c} \boldsymbol{\alpha}_z = 1$$

$$f(\boldsymbol{P}_c) = \max_{\boldsymbol{P}_c} \frac{\sum_{z=1}^{n_c} SSIM_{\boldsymbol{L}_{\boldsymbol{P}_c}, \boldsymbol{L}_{k_c(z)}}}{n_c}$$

end for
Generate final feature matrix $\boldsymbol{Q}$ using PSO with function $f(\boldsymbol{Q})$, $\boldsymbol{L}_{\boldsymbol{Q}}$ is the clustering matrix of $\boldsymbol{Q}$ with MLRSSC

$$\boldsymbol{Q} = \sum_{c=1}^{j} \boldsymbol{\beta}_c \boldsymbol{P}_c \;\; s.t. \;\; \sum_{c=1}^{j} \boldsymbol{\beta}_c = 1$$

$$f(\boldsymbol{Q}) = \max_{\boldsymbol{Q}} \frac{\sum_{c=1}^{j} SSIM_{\boldsymbol{L}_{\boldsymbol{P}_c}, \boldsymbol{L}_{\boldsymbol{Q}}}}{j}$$

The clustering result obtained according to $\boldsymbol{L}_{\boldsymbol{Q}}$ is the group surface partition template

The MLRSSC can only cluster brain surface common regions on a single subject. In order to obtain more effective clustering results in the brain, the Two-stage MLRSSC is proposed to achieve clustering at the group level. In the first stage, 100 subjects are clustered individually, and each subject is given a surface partition result. In the second stage, in order to better combine the information of the different objects, we first divided 100 individuals into j ($j = 10$) sets according to the similarity of all subjects' surface partition results. Then, set-inner surface partition results are generated by combining individuals of the same set with group clustering. Finally, the surface partitioning template on the groups is generated by combining all the set-inner surface partition results. The whole process of Two-stage MLRSSC is described as Algorithm.1.

For the similarity of subject surface partition results, SSIM was used to express it. After MLRSSC of 150 surface DRs of the subject, the clustering matrix $L_{150 \times 150}$ is obtained. $L_{i,j} = 1$ if DR i and j belong to the same cluster, otherwise $L_{i,j} = 0$. Then, the similarity $SSIM_{m,n}$ of L_m and L_n of two subject clustering matrices is Eq. (2).

$$\mathrm{SSIM}_{m,n} = \frac{\left(2\mu_{L_m}\mu_{L_n} + c_1\right)\left(2\sigma_{L_m L_n} + c_2\right)}{\left(\mu_{L_m}^2 + \mu_{L_n}^2 + c_1\right)\left(\sigma_{L_m}^2 + \sigma_{L_n}^2 + c_2\right)} \tag{2}$$

where μ_{L_m} is the average of L_m, and μ_{L_n} is the average of L_n. $\sigma_{L_m}^2$ is the variance of L_m, and $\sigma_{L_n}^2$ is the variance of L_n. $\sigma_{L_m L_n}$ is the covariance of L_m and L_n. Parameter c_1 and c_2 are used to keep the algorithm stable.

2.5 Exploring the Relationship Between Brain Function and Structure

In this work, the proposed approach is designed to explore the relationship between brain function and structure. We plan to investigate the relationship in each cluster, which are called common brain regions (CRs), obtained by Sect. 2.4. Each CR includes several brain DRs, so we calculate functional consistency matrix (FCM) and structural consistency matrix (SCM) among those brain DRs. FCM and SCM are obtained by calculating the correlation of the functional features and structural features, respectively. Then the pattern of those two matrices is further measured to study the relationship between brain function and structure.

2.6 Identifying the Brain Hierarchical Characteristics

In this work, another possible application of the proposed approach is to investigate the hierarchical characteristics based on brain function and structure. Since the proposed method can cluster the cortical surface into CRs, it is convenient to study the brain hierarchical characteristics from the obtained CRs by different scales (such as 10 CRs vs. 20 CRs). Therefore, we design different scales of brain CRs and to investigate the hierarchical characteristics based on the CRs obtained from different scales.

3 Results

3.1 Clustering Results and Relationship Between Function and Structure

By adopting the proposed method, a group surface partition template can be finally obtained, CRs_10 and CRs_20 results are provided separately in Fig. 3(A). We chose the number of brain surface clusters as 10 and 20, which benefits for exploring the brain hierarchical characteristics. The SSIM of the final template achieved 65.73% and 61.56% for CRs_10 and CRs_20. Compared with clustering the group features obtained by averaging of all individual features, our proposed method improves by 7.98% and 3.85%.

The functional and structural relationship within each CR in CRs_10 is further studied and results are shown in Fig. 3(B). By observing the SCM and FCM for each CR and analyzing the corresponding distribution of the types of their relationship, we found that the proposed method can effectively identify the common regions for brain function and structure. Specifically, two different categories can be clearly observed. CR 2, 5, 7 and 8 are the regions where brain function and structure are more complex and harder to identify the relationship, because their SCM and FCM are with low correlation. Those regions are more related to the high-level brain functions, e.g., the frontal lobe with working memory areas. On the contrary, CR 1, 3, 4, 6 and 9 are the regions where brain function and structure are relatively simple and stable, low-level brain functions are observed, e.g., the visual areas. More information can be obtained by the results of CRs_20, which we will study in future work.

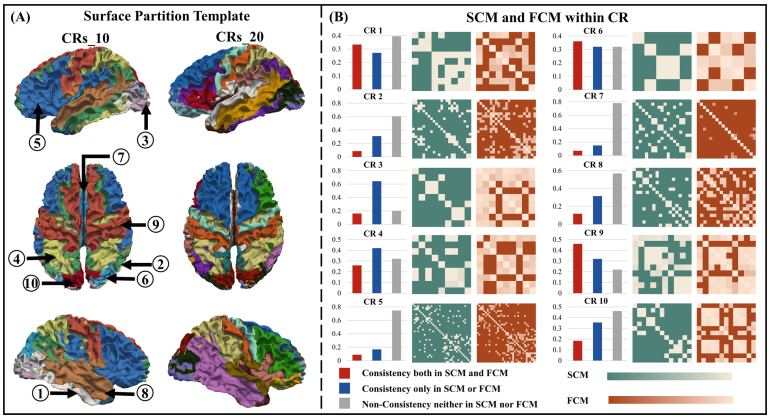


Fig. 3. Group surface partition template and the relationship of structure and function in CRs_10. (A) shows the group surface partition template of CRs_10 and CRs_20. (B) shows the SCM and FCM within each CR in CRs_10. The green matrix represents the SCM, and yellow matrix represents the FCM. The bar chart represents the consistency strength within each CR.

3.2 Further Analysis of the Cortical Partition Results

To further demonstrate the effectiveness of our proposed method, we compared the surface partition between the proposed Two-stage MLRSSC method and the gradient template in cortical features[11], as shown in Fig. 4. The gradient template represents how the spectrum of cortical function emerges from structural constraints. It is interesting that similar partitions can be identified from both templates, which are highlighted by colorful circles. This indicates that although different analyses of algorithms are designed, functional and structural relationship can be roughly identified.

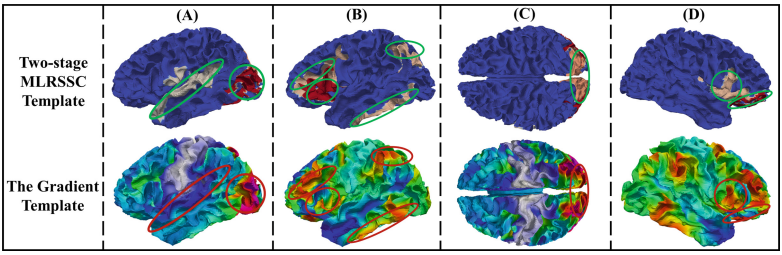


Fig. 4. Comparison of surface partition template with the gradient template.

3.3 Hierarchical Analysis on the Clustering Results

By proposed approach, different scales of the CRs can be easily figured out. It gives us a way to pay attention to their hierarchical characteristics. In Fig. 5(A), one of the CRs at high scale is divided into two or more CRs at low scale. In Fig. 5(B), one of the CRs in

high scale remains a CR in low scale. In Fig. 5(C), a portion of brain surface is divided into two CRs at high scale and into other two different CRs at low scale.

By comprehensively comparing the results of CRs_10 and CRs_20, an interesting phenomenon is observed. That is, for the simple and stable regions we mentioned above, e.g., CR 4 and 9 in high scale, they are still quite stable in low scale. However, for those complex regions, they are divided irregularly into several small pieces when the scales become finer. This finding reveals that more functions are gathered in those complex regions and finer scale representation are necessary.

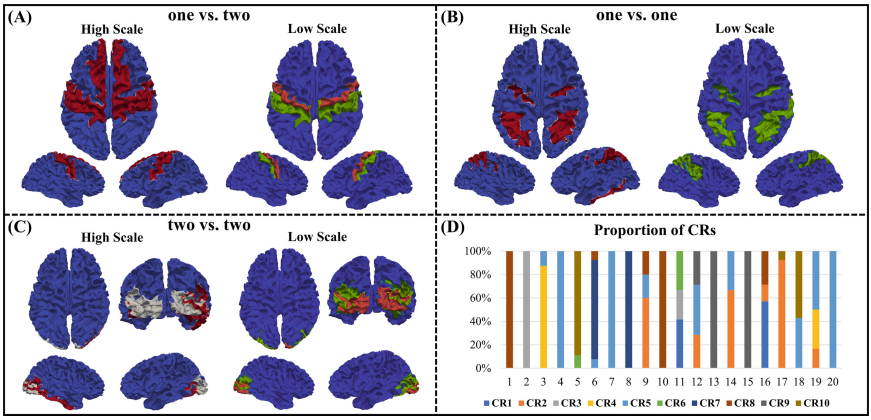


Fig. 5. Comparison between different scales of brain surface partition. We consider the CRs_10 and CRs_20 of surface partition templates as high and low scale templates, respectively. In (A), (B), and (C), we show three different results between CRs at different scales. (D) represents the proportion of CRs of at high scale in each CR at low scale.

4 Conclusion

In this work, a novel Two-stage MLRSSC method is proposed to jointly study the relationship between brain function and structure. The major innovation of the proposed approach is about the MLRSSC architecture, low-rank sparse representation can be obtained from functional and structural views to effectively study the relationship between them, then proposed two-stage strategy further improved the clustering performance and finally obtained the surface partition templates. The results are successfully evaluated when compared with related references. In addition, the proposed method can be utilized to further excavate and reveal the complexity of the brain function and structure, e.g., their hierarchical characteristics. Overall, the proposed Two-stage MLRSSC method provides a new way to study and explore the relationship between brain function and structure. Much more detailed studies will be settled in future work.

References

1. Sporns, O., Tononi, G., Kötter, R.: The human connectome: a structural description of the human brain. PLoS Comput. Biol. **1**, e42 (2005)

2. Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S.: Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn. Reson. Med. **34**, 537–541 (1995)
3. Gordon, E.M., et al.: Precision functional mapping of individual human brains. Neuron **95**, 791-807.e7 (2017)
4. Zamora-López, G., Zhou, C., Kurths, J.: Cortical hubs form a module for multisensory integration on top of the hierarchy of cortical networks. Front. Neuroinform. **4**, 1 (2010)
5. Sui, J., Adali, T., Yu, Q., Chen, J., Calhoun, V.D.: A review of multivariate methods for multimodal fusion of brain imaging data. J. Neurosci. Methods **204**, 68–81 (2012)
6. Stam, C.J., et al.: The relation between structural and functional connectivity patterns in complex brain networks. Int. J. Psychophysiol. **103**, 149–160 (2016)
7. Liang, H., Wang, H.: Structure-function network mapping and its assessment via persistent homology. PLoS Comput. Biol. **13**, e1005325 (2017)
8. Meier, J., et al.: A mapping between structural and functional brain networks. Brain Connect. **6**, 298–311 (2016)
9. Smith, S.M., et al.: Functional connectomics from resting-state fMRI. Trends Cogn. Sci. **17**, 666–682 (2013)
10. Brbić, M., Kopriva, I.: Multi-view low-rank sparse subspace clustering. Pattern Recogn. **73**, 247–258 (2018)
11. Huntenburg, J.M., Bazin, P.-L., Margulies, D.S.: Large-scale gradients in human cortical organization. Trends Cogn. Sci. **22**, 21–31 (2018)
12. Destrieux, C., Fischl, B., Dale, A., Halgren, E.: Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage **53**, 1–15 (2010)
13. Van Essen, D.C., et al.: The human connectome project: a data acquisition perspective. Neuroimage **62**, 2222–2231 (2012)
14. Dale, A.M., Fischl, B., Sereno, M.I.: Cortical surface-based analysis: I segmentation and surface reconstruction. Neuroimage **9**, 179–194 (1999)
15. Fischl, B., et al.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron **33**, 341–355 (2002)
16. Glasser, M.F., et al.: The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage **80**, 105–124 (2013)
17. van den Heuvel, M.P., Kahn, R.S., Goñi, J., Sporns, O.: High-cost, high-capacity backbone for global brain communication. Proc. Natl. Acad. Sci. **109**, 11372–11377 (2012)
18. Jenkinson, M., Bannister, P., Brady, M., Smith, S.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage **17**, 825–841 (2002)
19. Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M.: FSL. Neuroimage **62**, 782–790 (2012)
20. Zhu, D., et al.: DICCCOL: dense individualized and common connectivity-based cortical landmarks. Cereb. Cortex **23**, 786–800 (2013)
21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**, 600–612 (2004)
22. Kennedy, J., Eberhart, R.C.: A discrete binary version of the particle swarm algorithm. In: Computational Cybernetics and Simulation 1997 IEEE International Conference on Systems, Man, and Cybernetics, vol. 5, pp. 4104–4108 (1997)

Fast Image-Level MRI Harmonization via Spectrum Analysis

Hao Guan[1], Siyuan Liu[2], Weili Lin[1], Pew-Thian Yap[1], and Mingxia Liu[1(✉)]

[1] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
mingxia_liu@med.unc.edu

[2] Marine Engineering College, Dalian Maritime University, Dalian 116026, China

Abstract. Pooling structural magnetic resonance imaging (MRI) data from different imaging sites helps increase sample size to facilitate machine learning based neuroimage analysis, but usually suffers from significant cross-site and/or cross-scanner data heterogeneity. Existing studies often focus on reducing cross-site and/or cross-scanner heterogeneity at handcrafted feature level targeting specific tasks (e.g., classification or segmentation), limiting their adaptability in clinical practice. Research on image-level MRI harmonization targeting a broad range of applications is very limited. In this paper, we develop a spectrum swapping based image-level MRI harmonization (SSIMH) framework. Different from previous work, our method focuses on alleviating cross-scanner heterogeneity at *raw image level*. We first construct *spectrum analysis* to explore the influences of different frequency components on MRI harmonization. We then utilize a *spectrum swapping* method for the harmonization of raw MRIs acquired by different scanners. Our method does not rely on complex model training, and can be directly applied to fast real-time MRI harmonization. Experimental results on T1- and T2-weighted MRIs of phantom subjects acquired by using different scanners from the public ABCD dataset suggest the effectiveness of our method in structural MRI harmonization at the image level.

Keywords: MRI · Image-level harmonization · Spectrum analysis

1 Introduction

Structural magnetic resonance imaging (MRI) data acquired from different sites or scanners may have significant differences, which has become one of the major concerns of the medical imaging community. Those between-site or between-scanner variabilities may negatively influence the analysis results of multi-site/multi-center studies and set barrier to building large-scale publicly available medical image datasets. As for computer-aided diagnosis, since conventional machine learning theory typically assumes the training dataset and test dataset follow the same distribution, thus the distribution variability among

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 201–209, 2022.
https://doi.org/10.1007/978-3-031-21014-3_21

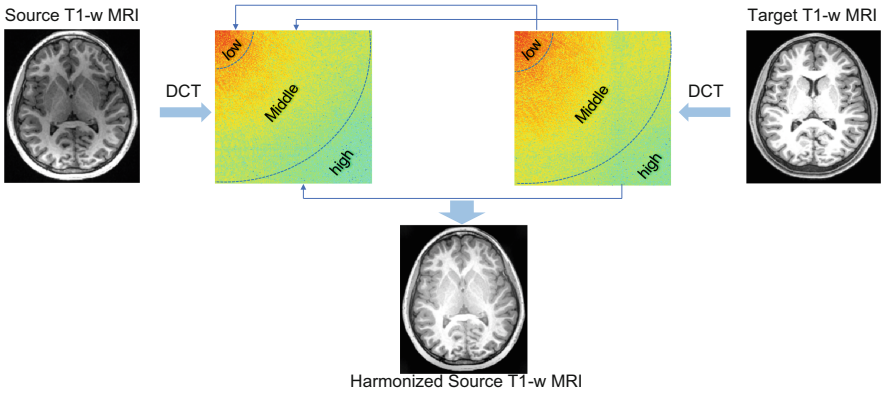


Fig. 1. Illustration of the spectrum swapping based image-level MRI harmonization (SSIMH) framework via spectrum analysis. Discrete Cosine Transform (DCT) is facilitated to source and target MRIs to project them into frequency domain. The frequency domain is divided into three spectra: low-frequency, middle-frequency, and high-frequency regions. Through swapping the key domain-variant frequency regions, harmonization can be achieved.

different sites/scanners would degrade the performance of machine learning systems. A promising solution to handle the between-site/scanner variability is MR image harmonization. To this end, domain adaptation or transfer learning techniques have been extensively introduced into medical image harmonization [1–3]. However, domain adaptation-based harmonization methods generally work at the feature level which is typically associated with specific tasks, e.g., classification or segmentation [4,5]. These methods remove dataset biases in some learning modules (e.g., networks) without constructing harmonized MRIs at the whole-image level. A more general solution is *image-level MRI harmonization* which can be used as a universal pre-processing step to remove site or scanner bias without considering any specific downstream tasks. These methods fall into two categories: 1) non-training methods, such as intensity histogram matching (HM) [6]; and 2) deep learning methods, such as GAN [7–11]. Since only overall intensity distribution is considered without accounting for structure information, the intensity histogram matching methods may lose some detailed information within MR images. For those deep learning methods, they usually need a relatively large amount of images for complex and time-consuming model training. Also, for images from new domains/sites, they often need model re-training or fine-tuning for harmonization. These shortcomings will significantly limit their utility and flexibility in real-world applications.

In this paper, motivated by [12,13], we propose a spectrum swapping based image-level MRI harmonization (SSIMH) through spectrum analysis. Different from previous studies, our analysis and harmonization method works in the frequency domain using Discrete Cosine Transform (DCT) as the primary analysis tool. As shown in Fig. 1, an MRI is first transformed into the frequency

domain by DCT. We divide the spectrum of each MRI into three frequency areas: low-frequency, middle-frequency (further divided into several sub-regions) and high-frequency. We first explore an essential problem: which is the most influential frequency region for image-level MRI harmonization. Then by swapping the corresponding frequency region of the source and target image, image-level harmonization on the raw MRIs is achieved. Our framework not only avoids complex data-driven training, but also achieves image level MRI harmonization which can offer more flexibility for clinical usage. We use T1-weighted (T1-w) and T2-weighted (T2-w) MRIs of traveling phantom MRIs acquired by using different scanners from the ABCD dataset [14] to verify the effectiveness of our methods and its ability for harmonization of multi-modality data.

2 Materials and Experimental Settings

Materials. We use the public ABCD dataset [14] with structural MRIs for analysis[1]. The ABCD dataset includes brain MRIs of nearly 12,000 youth for the research of brain development, mental health and other related topics. We use T1-w and T2-w MR images in this dataset for MRI harmonization evaluation. Five traveling phantom subjects with raw T1-weighted MRIs, and five phantom subjects with raw T2-weighted MRIs are adopted. We select three slices in the middle part of each MRI (in axial view) for harmonization. The MRIs of each traveling phantom subject were acquired by three different scanners, include GE scanner, Philips scanner and Siemens scanner. We do not perform standard brain MRI pre-processing (e.g., skull stripping or registration) and only use raw MRIs as the input of our method and each competing method.

Evaluation Metrics. The traveling phantom data from the ABCD dataset which are acquired by using three scanners are used to assess the harmonization capability of different methods. Given a traveling phantom subject scanned at scanners A and B, we have two MR images M_A and M_B accordingly. Let M_B be the target image (reference image). Then, M_A is harmonized to M_B through certain methods, and we get the harmonized MR image M'_A. For an evaluation metric $\mathcal{L}(\cdot)$, we calculate the value of $\mathcal{L}(M_A, M_B)$ and $\mathcal{L}(M'_A, M_B)$, respectively, and compare them to check the harmonization performance. Please note that $\mathcal{L}(M_A, M_B)$ is the baseline performance without any harmonization. We use the following metrics to evaluate the harmonization performance.

- **Image Similarity.** We calculate 2D correlation coefficient (CC) between M'_A and M_B. Larger CC values indicate better performance.
- **Peak Signal-to-Noise Ratio (PSNR)**. We calculate the peak signal-to-noise ratio (PSNR) between M'_A and M_B. Larger PSNR values indicate better results.
- **Mean-Squared Error (MSE)**. We calculate the mean-squared error (MSE) between M'_A and M_B. Smaller MSE values indicate better performance.

[1] https://nda.nih.gov/abcd.

– **Visual Inspection.** We also use visual inspection as a qualitative evaluation metric to evaluate the harmonization results. The original source image, harmonized source image, and the target image are visually compared in the spatial space for inspection.

Competing Methods. We compare the proposed SSIMH method with two approaches for harmonization performance analysis in our experiment, including: (1) Baseline without harmonization. The raw phantom MRIs acquired by two different scanners are directly compared using the evaluation metrics without any harmonization. (2) Intensity histogram matching (HM) [6]. The intensity histogram of the source phantom MRI is matched with the intensity histogram of the target phantom MRI to make their intensity distribution more similar. It should be noted that we avoid using the complex model-based methods (e.g., GAN) because they need a large amount of MRI data and much training time.

3 Spectrum Analysis for Image-Level Harmonization

Discrete Cosine Transform Analysis. The Discrete Cosine Transform (DCT) is a spectrum analysis tool which can be applied to image pixels in the spatial domain and transform them into a frequency domain. The DCT is commonly used in image compression with the advantage of information concentration and avoidance of complex number computation. Specifically, for an input image $\mathbf{I}_{M\times N}$, the corresponding spectrum image via DCT is a matrix $\mathbf{T}$ with each entity $\mathbf{T}(i,j)$ calculated as:

$$\mathbf{T}(i,j) = \alpha_i \alpha_j \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \mathbf{I}(m,n) \cos\frac{\pi(2m+1)i}{2M} \cos\frac{\pi(2n+1)j}{2N}$$

$$\alpha_i = \begin{cases} \frac{1}{\sqrt{M}}, & \text{i} = 0 \\ \sqrt{\frac{2}{M}}, & 1 \leq \text{i} \leq \text{M-1} \end{cases}$$

$$\alpha_j = \begin{cases} \frac{1}{\sqrt{N}}, & \text{j} = 0 \\ \sqrt{\frac{2}{N}}, & 1 \leq \text{j} \leq \text{N-1} \end{cases}$$

Spectrum Analysis. We conduct spectrum analysis for MRI harmonization. Specifically, a pair of phantom MRIs (a source MRI and a target MRI) are firstly transformed into frequency domain through DCT. Then we use five different spectrum filters to filter out different frequency components, i.e., low-frequency part, middle-frequency part (further divided into three sub-regions), and high-frequency part. As shown in Fig 2, for a $M \times M$ image (slice), we get a $M \times M$ DCT image (in frequency domain). In our experiment, M is set to 250. Then the radius of the low-frequency part is 0∼5, while the radii for the three sub-regions of middle-frequency are 5∼20, 20∼100, 100∼250, respectively. The left region is the high-frequency part. For each time, we replace one frequency part of the

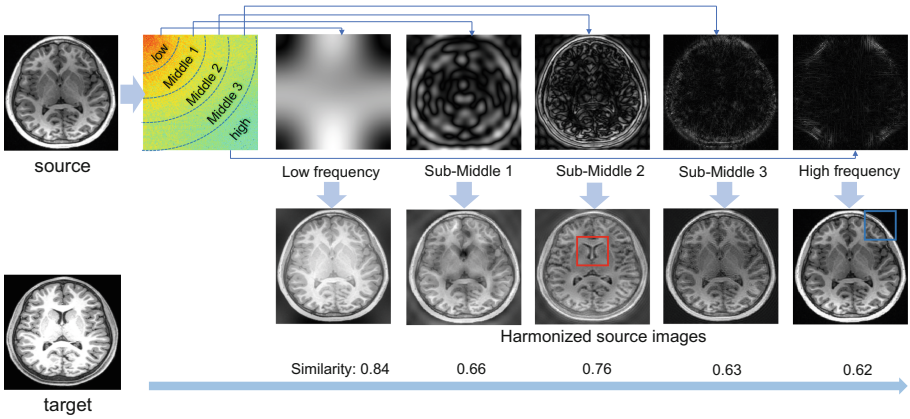


Fig. 2. Influence of different spectrum/frequency of MRIs on harmonization performance.

source image with the corresponding frequency part of the target image. Then, we use inverted DCT to transform the spectrum source image back to the spatial domain, and test the corresponding harmonization performance. The analysis results are shown in Fig. 2. From the results, we have the following findings. (1) *Low-frequency part of MRI has the largest influence on the harmonization.* For example, from Fig. 2, through replacing the low-frequency part of source image with the corresponding low-frequency region of target image, the harmonized source image achieves the highest similarity with the target image, and there is no structure corruption from the original source image. (2) *Detailed structure information mainly lies in the middle-frequency region.* As shown in the red box in Fig. 2, after replacing the second sub middle-frequency part of source image with corresponding region in target image (the 4th column), the harmonized source image has corrupted into the target domain without preserving the original anatomical structure. This indicates the anatomical structure information of MRI is mainly reflected by the middle-frequency region. (3) *High-frequency region has limited contribution to harmonization.* From Fig. 2, the high-frequency region has very small values which can only reflect some edges of the MRIs (see the blue box). Swapping these regions between source and target image has barely effects on the harmonization performance. It should be noted that in real-world practice, there are often noises in the MR images. Thus image quality check (QC) and image denoising need to be performed.

4 Spectrum Swapping for Image-Level Harmonization

Harmonization on Phantom MRIs. Through the spectrum analysis, we identify that the low-frequency part mainly influences harmonization. Thus we

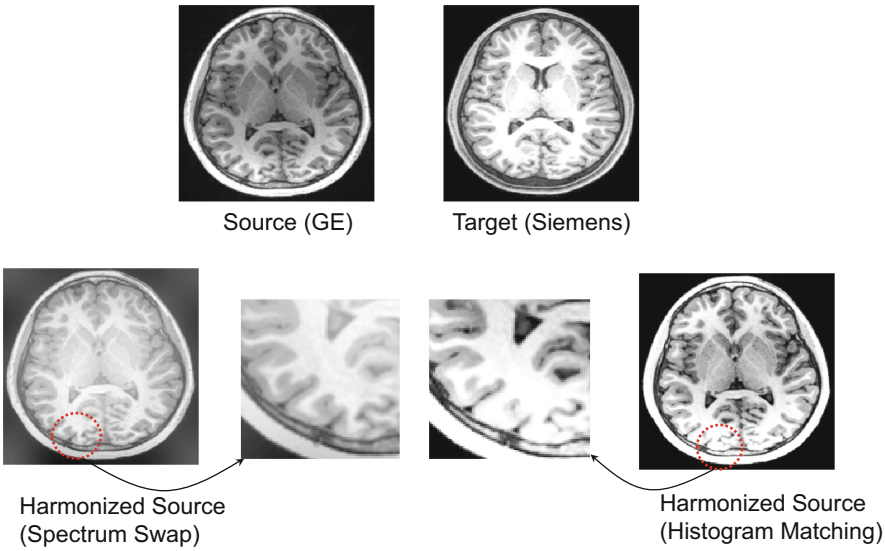


Fig. 3. Visual inspection of MRI harmonization results achieved by two methods.

Table 1. Harmonization of T1- and T2-weighted MRIs of five traveling phantom subjects acquired by three different scanners in the ABCD dataset.

Source→Target	Method	T1-weighted MRI			T2-weighted MRI		
		CC	PSNR	MSE	CC	PSNR	MSE
GE→Siemens	Baseline	0.59±0.06	9.61±0.29	0.11±0.01	0.73±0.06	11.66±0.82	0.07±0.01
	HM	0.67±0.06	10.45±0.60	0.09±0.01	0.75±0.07	12.26±1.38	0.06±0.02
	SSIMH	**0.84±0.02**	**13.25±0.73**	**0.05±0.01**	**0.81±0.03**	**13.37±0.75**	**0.05±0.01**
Siemens→Philips	Baseline	0.73±0.03	11.78±0.45	0.06±0.01	0.81±0.01	12.79±0.19	0.05±0.001
	HM	0.74±0.03	12.00±0.47	0.06±0.01	0.82±0.002	12.69±0.05	0.05±0.002
	SSIMH	**0.76±0.02**	**12.81±0.38**	**0.05±0.002**	**0.86±0.01**	**17.60±0.35**	**0.04±0.002**
Philips→GE	Baseline	0.64±0.02	10.70±0.17	0.09±0.003	0.75±0.02	11.31±0.44	0.07±0.01
	HM	0.62±0.03	11.25±0.11	0.08±0.002	0.68±0.03	11.02±0.45	0.08±0.01
	SSIMH	**0.70±0.03**	**12.83±0.23**	**0.05±0.002**	**0.85±0.03**	**14.43±0.82**	**0.04±0.01**

replace the low-frequency region of the source image with the corresponding low-frequency region of target image, and project the spectrum source image back into the spatial domain. We conduct harmonization on MRIs from five traveling phantom subjects in the ABCD dataset. Each phantom is acquired by two different scanners, *i.e.*, GE scanner, Siemens scanner or Philips scanner. We use these phantoms to test the performance of different methods for cross-scanner MRI harmonization. The results in terms of three metrics are shown in Table 1. From Table 1, it can be seen that the proposed method consistently outperforms the baseline and the HM methods by significant margins. In addition, since no data-based training is required, the running speed of the proposed method is also fast, *i.e.*, within 1 s on a Intel i-7 CPU with 16 GB memory.

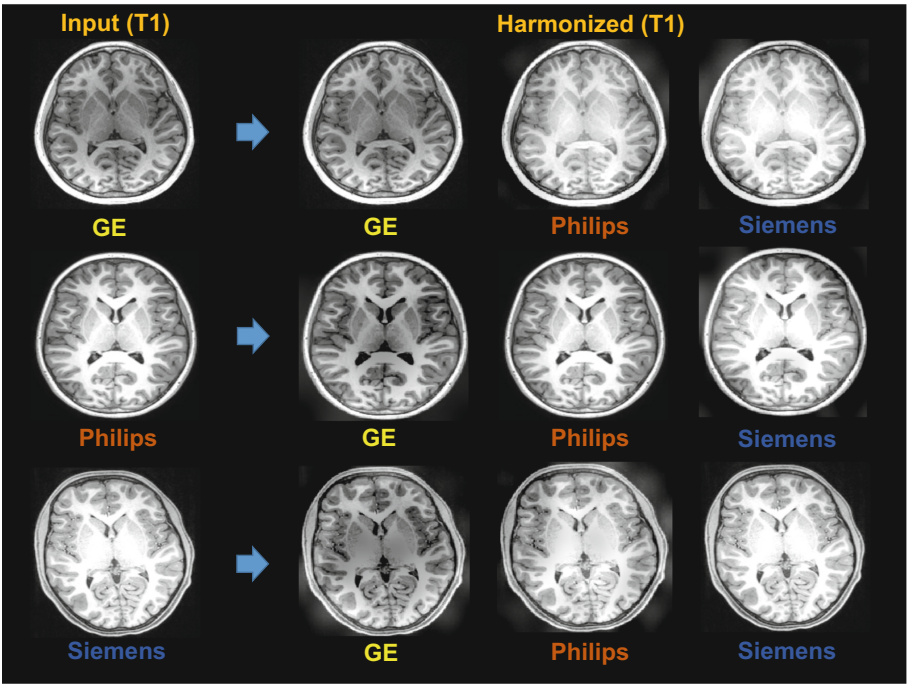


Fig. 4. Visualization of harmonization results of our SSIMH method for T1-w MRIs acquired by different scanners (*i.e.*, GE, Philips, Siemens).

We also utilize visual inspection to explore the harmonization performance. A visualization result is shown in Fig. 3. The phantom is acquired by a GE scanner (source) and Siemens scanner (target), respectively. After harmonization, it can be seen that the histogram matching method loses some detailed information such as some gyrus areas in cortex (see red dotted circles), while our SSIMH method can well preserve these original anatomical structures.

Generic Cross-Scanner Harmonization Without Phantom. Besides using phantom data for harmonization evaluation, we also select independent subjects that acquired by using different scanners, *i.e.*, GE, Philips, and Siemens. Each subject is harmonized to the MRI styles of the other two scanners. We conduct harmonization on two groups of MRIs. One is T1-w images, while the other group is T2-w images. The visualization of the harmonization for T1-w and T2-w MRIs are shown in Figs. 4 and 5, respectively. From the results, it can be seen that the MRIs acquired with different scanners have significantly different imaging styles, and the proposed method can harmonize MRIs to different scanner styles well. Since the proposed method does not need pre-training using numerous MRIs or phantom data, it is very flexible and convenient to use in practice.

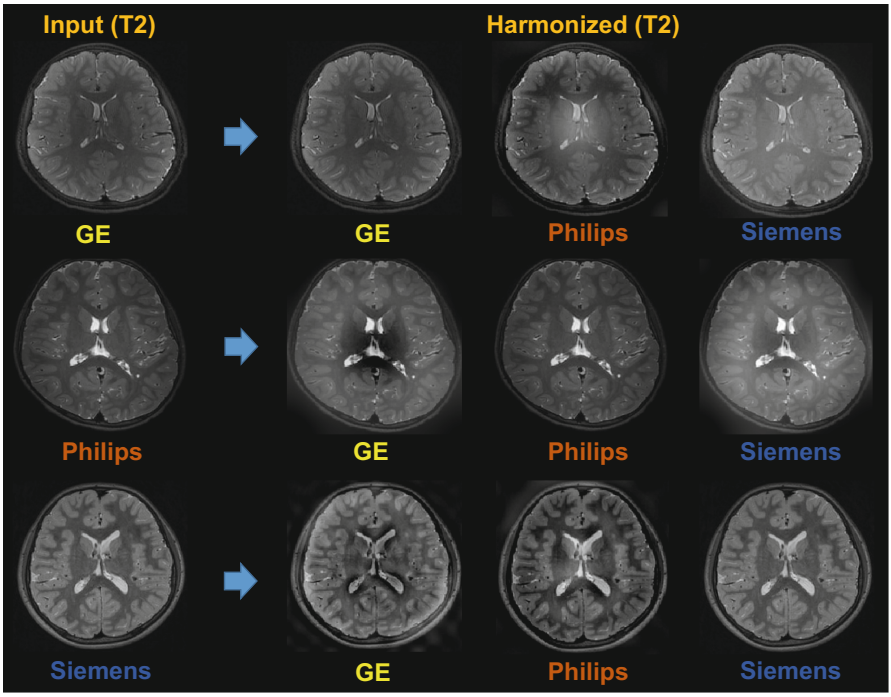


Fig. 5. Visualization of harmonization results of our SSIMH method for T2-w MRIs acquired by different scanners (*i.e.*, GE, Philips, Siemens).

5 Conclusion

We present a spectrum swapping based image-level MRI harmonization (SSIMH) framework based on spectrum analysis. The Discrete Cosine Transform (DCT) is used as the tool to project the source and target images into the frequency domain. We divide the frequency domain into low-frequency, middle-frequency and high-frequency regions. We firstly analyze the influence of each region on MRI harmonization, and find that the low-frequency region plays a more important role in effective harmonization, while the middle-frequency helps keep the detailed structure information in MRIs. Through replacing the low-frequency region of source images with the corresponding low-frequency region of target image, we conduct harmonization of MRIs acquired by using different scanners. Experiments on T1- and T2-weighted MRIs from the ABCD dataset suggest that the proposed method achieves satisfactory image-level harmonization results. Another advantage of the spectrum-based method is the avoidance of complex training which relies on a large amount of training data and is time-consuming.

Acknowledgement. This work was partly supported by NIH grants (Nos. AG073297 and AG041721).

References

1. Cheng, B., Liu, M., Zhang, D., Shen, D.: Robust multi-label transfer feature learning for early diagnosis of Alzheimer's disease. Brain Imaging Behav. **13**(1), 138–153 (2019)
2. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. **54**, 280–296 (2019)
3. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. **69**(3), 1173–1185 (2022)
4. Dinsdale, N.K., Jenkinson, M., Namburete, A.I.: Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. Neuroimage **228**, 117689 (2021)
5. Guan, H., et al.: Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. Med. Image Anal. **71**, 102076 (2021)
6. Shinohara, R.T., et al.: Statistical normalization techniques for magnetic resonance imaging. NeuroImage. Clin. **6**, 9–19 (2014)
7. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
8. Zuo, L., et al.: Information-based disentangled representation learning for unsupervised MR harmonization. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (eds.) IPMI 2021. LNCS, vol. 12729, pp. 346–359. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78191-0_27
9. Dewey, B.E., et al.: Deepharmony: a deep learning approach to contrast harmonization across scanner changes. Magn. Reson. Imaging **64**, 160–170 (2019)
10. Sinha, S., Thomopoulos, S.I., Lam, P., Muir, A., Thompson, P.M.: Alzheimer's disease classification accuracy is improved by MRI harmonization based on attention-guided generative adversarial networks. In: 17th International Symposium on Medical Information Processing and Analysis, vol. 12088, pp. 180–189. SPIE (2021)
11. Cackowski, S., Barbier, E.L., Dojat, M., Christen, T.: Imunity: a generalizable VAE-GAN solution for multicenter MR image harmonization. arXiv preprint arXiv:2109.06756 (2021)
12. Huang, J., Guan, D., Xiao, A., Lu, S.: FSDR: Frequency space domain randomization for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6891–6902 (2021)
13. Yang, Y., Soatto, S.: FDA: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4085–4095 (2020)
14. Volkow, N.D., et al.: The conception of the ABCD study: from substance use to a broad NIH collaboration. Dev. Cogn. Neurosci. **32**, 4–7 (2018)



CT2CXR: CT-based CXR Synthesis for Covid-19 Pneumonia Classification

Peter Ho Hin Yuen[1], Xiaohong Wang[2], Zhiping Lin[1], Nikki Ka Wai Chow[3], Jun Cheng[2], Cher Heng Tan[4], and Weimin Huang[2(✉)]

[1] School of EEE, Nanyang Technological University, Singapore, Singapore
[2] Institute for Infocomm Research, A*Star, Singapore, Singapore
wmhuang@i2r.a-star.edu.sg
[3] The University of Hong Kong, Hong Kong, China
[4] Tan Tock Seng Hospital, Singapore, Singapore

Abstract. Chest X-ray (CXR) is a common imaging modality for examination of pneumonia. However, some pneumonia signs which are visible in CT may not be clearly identifiable in CXR. It is challenging to create a good ground truth for positive pneumonia cases based on CXR images especially for cases with small pneumonia lesions. In this paper, we propose a novel CT-based CXR synthesis framework, called ct2cxr, to perform data augmentation for pneumonia classification. Generative Adversarial Networks (GANs) were exploited and a customized loss function was proposed for model training to preserve the target pathology and maintain high image fidelity. Our results show that CXR images generated through style mixing can enhance the performance of general pneumonia classification models. Testing the models on a Covid-19 dataset shows similar improvements over the baseline models.

Keywords: Generative adversarial network · Data augmentation · Covid-19 · Pneumonia · CXR Synthesis

1 Introduction

Since the onset of the Covid-19 pandemic, there has been a global effort in developing deep learning-based diagnostic models. While Polymerase Chain Reaction (PCR) tests have become the gold standard to detect Covid-19 infections, radiological investigations are often necessary when it comes to the diagnosis of Covid-19 pneumonia. Various models have been proposed for fast and reliable diagnosis of Covid-19 pneumonia based on chest X-ray (CXR) images [1,5,10,22] and CT scans [14,28,29] respectively. At the same time, concerns have been raised about the generalizability of such models due to the inconsistent quality of available datasets [16].

Usually, in a radiological investigation, CXR images are taken as a first step before further diagnosis with more in-depth imaging modalities such as CT scans as CXR images are more affordable and easy to obtain. However, some pneumonia signs which are visible in CT may not be clearly identifiable in CXR,

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 210–219, 2022.
https://doi.org/10.1007/978-3-031-21014-3_22

and hence it is challenging to create reliable ground truth for positive pneumonia cases based on CXR images when cases with small pneumonia lesions are included for consideration. In light of significant proportion of false negative pneumonia based on CXR datasets, CT-based data augmentation for training deep learning models is important and timely. Developing such a model will allow for an accurate depiction of disease at baseline, while creating opportunities for follow-up with CXR imaging rather than repeated CT scans (that contain higher doses of ionising radiation). CXRs are also much quicker to assess in clinical practice, compared to multislice CT images.

One way to achieve CT-based data augmentation is through CT-CXR conversions which are usually performed using Digitally Reconstructed Radiograph (DRR) techniques [11,20,23]. DRR methodologies rely on simulation of physical effects such as scattering and noise injection to generate a projected CXR view. A recent paper introduced a technique called DeepDRR which uses deep learning to perform material segmentation and scatter estimation but still performs forward projection analytically [25]. The main advantage of DRR algorithms lies in the fact that they can generate CXR views at arbitrary angles. However, in CXR data augmentation use cases, we are mostly concerned with frontal CXR views. The weakness of existing DRR algorithms is also obvious, being that the resolution of the generated CXR images completely depends on the slice thickness and resolution of the original CT scans.

In this paper, we propose a novel CT-based data augmentation framework, called ct2cxr, to perform CXR synthesis for pneumonia classification. We exploit Generative Adversarial Networks (GANs) in our framework. GAN, first proposed in 2014 [6], is a deep learning technique suitable for generative tasks. In recent years, GANs have found many successes in tasks such as photo-realistic image generation, super-resolution and image-to-image translation [27]. While there are recent works on CXR synthesis from noise for Covid-19 pneumonia classification [13,26], to our best knowledge, there is no prior work in GAN-based CT-CXR translation and simulation. In our proposed CT-based CXR synthesis framework, we use GAN inversion techniques, which can achieve controllable image generation given any form of inputs, and have been applied in some medical image generation such as in mammogram synthesis and segmentation-CT/MR translation [12,17]. Compared with DRR methods or pre-existing style-transfer GAN architectures such as CycleGAN [32], our framework is able to perform super-resolution and achieve cross-modal synthesis through style mixing.

The main contributions of our work are summarized as follows:

1. Application of GAN to the domain of CT-based CXR augmentation;
2. Proposing a novel loss function to encourage pathology preservation;
3. Enabling cross-modal synthesis through style mixing of multiple CT scans;
4. Validating the effectiveness of the proposed data augmentation framework in enhancing the performance of pneumonia classification models.

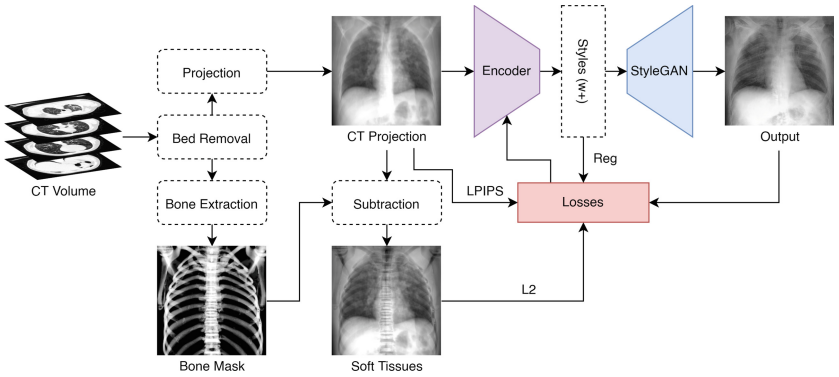


Fig. 1. The ct2cxr framework. This figure illustrates the CT-CXR synthesis process. The losses are used during the encoder training. The *Projection* step involves summing up the Hounsfield values along the frontal axis followed by min-max normalization. *Bed Removal* and *Bone Extraction* are performed using traditional computer vision methods such as thresholding, contour-finding and dilation-erosion.

2 Methodology

Our proposed ct2cxr framework is largely inspired by pixel2style2pixel (pSp), an encoder-based image-to-image translation framework [18,24]. Training is performed in two stages. First, an unconditional StyleGAN is pretrained on a large-scale generic CXR dataset using the same procedures as the original StyleGAN2 paper [9]. Then, the encoder is trained using a CT dataset while the StyleGAN weights are fixed (see Fig. 1).

2.1 StyleGAN Pretraining

We pretrained our StyleGAN network using a subset of the CheXpert dataset [8]. One CXR image in the AP orientation was randomly chosen per patient and images with supporting devices were removed. There are multiple reasons for not using a Covid-19 pneumonia CXR dataset for the StyleGAN pretraining. Firstly, using a large scale dataset exposes the StyleGAN to a wider variety common CXR features such as bone structures and more prominent anatomical features (e.g. heart and vessels). Secondly, the training of StyleGAN is a time- and resource-consuming process. Having a general-purpose StyleGAN pretrained on a large scale dataset would allow other researchers with limited computing resources to adapt only the encoder for their specific use cases.

2.2 Encoder Training

The second stage of ct2cxr involves training an encoder network. The encoder architecture follows the pSp framework, using a feature pyramid with a ResNet backbone to extract feature maps from the input image. The feature maps are

subsequently mapped to style vectors and added to the mean style vector $\bar{\mathbf{w}}$ of the pretrained generator before being passed to the generator. Hence, given an input image $\mathbf{x}$, the output image is:

$$ct2cxr(\mathbf{x}) = G(E(\mathbf{x}) + \bar{\mathbf{w}}),$$

where $G(\cdot)$ denotes the pretrained generator and $E(\cdot)$ denotes the encoder.

We trained our encoder using the COVID19-CT-Dataset [15]. From each of the 1013 patients, we selected the scan (mediastinum window) with the lowest slice thickness. We manually removed scans with poor quality (with blurred image or strong artifact such as motion) and no identifiable COVID-19 pneumonia lesion in CT to obtain 642 training samples. The CT samples were then processed according to Fig. 1 to form the inputs to the encoder.

2.3 Loss Functions

We have performed various experiments to adapt the loss functions proposed in pSp to our use case. In ct2cxr, the $L2$ loss is modified and the identity loss is removed because identity preservation is not critical for data augmentation tasks. Our final loss function is:

$$\mathcal{L} = \lambda_{L2}\mathcal{L}_{L2} + \lambda_{LPIPS}\mathcal{L}_{LPIPS} + \lambda_{reg}\mathcal{L}_{reg},$$

$$\mathcal{L}_{L2}(\mathbf{x}) = ||soft(\mathbf{x}) - ct2cxr(\mathbf{x})||_2,$$

$$\mathcal{L}_{LPIPS}(\mathbf{x}) = ||F(\mathbf{x}) - F(ct2cxr(\mathbf{x}))||_2,$$

$$\mathcal{L}_{reg}(\mathbf{x}) = ||E(\mathbf{x}) - \bar{\mathbf{w}}||_2,$$

where $soft$ refers to soft tissues extraction and $F(\cdot)$ is a pretrained VGG-19 perceptual feature extractor. The coefficients were chosen to be $\lambda_{L2} = 3, \lambda_{LPIPS} = 0.1, \lambda_{reg} = 0.01$.

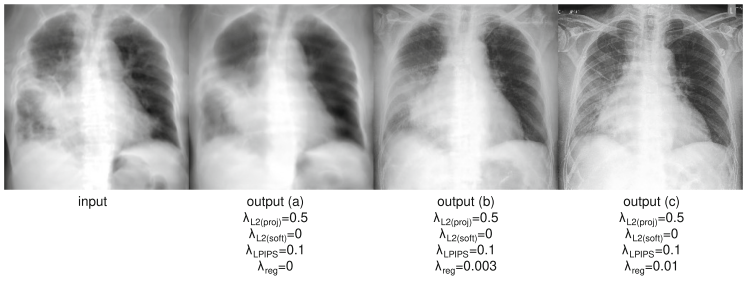


Fig. 2. Effect of the regularization loss.

Regularization Loss. We tested the effect of the regularization loss and found out that the regularization loss controls the trade-off between pathology preservation and image quality. We have trained 3 different models using different values of λ_{reg} to demonstrate the difference (Fig. 2). The difference can be understood intuitively as a high regularization loss coefficient restricts the output style vector to a smaller, high-fidelity region around the mean.

L2 Loss. Realizing that the regularization loss controls the quality and fidelity of the image, we no longer need to rely on the $L2$ loss to generate anatomical structures which do not need to be preserved. As long as an appropriate λ_{reg} is chosen, the output image would contain high-fidelity body structures. As such, we changed the ground truth of the $L2$ loss from the CT projection to the soft tissues image, which contains our target pathology: Covid-19 pneumonia lesions (i.e. from $||\mathbf{x} - ct2cxr(\mathbf{x})||_2$ to $||soft(\mathbf{x}) - ct2cxr(\mathbf{x})||_2$). Figure 3 demonstrates how the change in $L2$ loss results in better preservation of pneumonia features. This technique can be adapted to any other desired pathologies. Projection image of the target pathology can be first extracted from CT using segmentation methods and subsequently used in the $L2$ loss to encourage preservation of the pathology.

LPIPS Loss. Having modified the $L2$ loss, the $LPIPS$ loss [30] is necessary for the model to achieve maximum perceptual similarity between the output and input images with a solution space constrained by λ_{reg} and pathology preservation λ_{L2}.

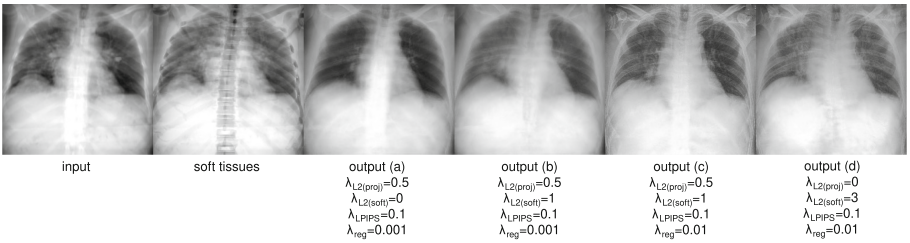


Fig. 3. Effect of changing $L2$ loss ground truth. Comparison between (a) and (b) shows how pneumonia signs in the right lung is better preserved using the soft tissues ground truth. Comparison between (b) - (d) shows the effects of $L2$ and regularization losses combined.

3 Experiments and Results

3.1 Latent Space Exploration

To ensure that our ct2cxr model is able to capture pneumonia pathology present in the dataset, we performed a latent space exploration. We trained

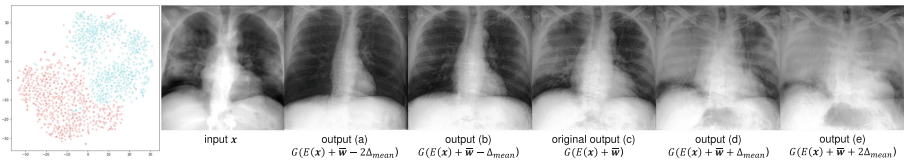


Fig. 4. Latent space exploration. The t-SNE visualization on the left shows the separation in style vectors between the Covid-19 pneumonia (red) and non-pneumonia datasets (cyan). The images show the effect of adding / subtracting the difference in mean latent vector. An increase in pneumonia severity can be seen from left to right. (Color figure online)

a second model using the LIDC-IDRI lung nodules dataset [2] for synthesis of non-pneumonia CXRs. We then calculated the difference in mean style vector between the pneumonia and non-pneumonia datasets:

$$\Delta_{mean} = \underset{\mathbf{x}_+ \in X_+}{avg}\ (E_+(\mathbf{x}_+)) - \underset{\mathbf{x}_- \in X_-}{avg}\ (E_-(\mathbf{x}_-)), \quad \Delta_{mean} \in \mathcal{W}+,$$

where X_+ and X_- are the pneumonia and non-pneumonia datasets. E_+ and E_- are the encoders trained on X_+ and X_- respectively.

Δ_{mean} should point in the direction of increasing Covid-19 pneumonia severity in the extended latent space $\mathcal{W}+$. Figure 4 confirms the result and shows that the ct2cxr model is able to learn a representation of pneumonia signs in the $\mathcal{W}+$ space of the pretrained generator.

3.2 Image Quality

Table 1. Image quality metrics of each synthesis method compared to BrixIA Covid-19 pneumonia CXR dataset [4,21]. (IS: higher is better, FID, KID: lower is better)

Synthesis method	IS	FID	KID
CT projection	**2.60**	123	0.132
DeepDRR	1.96	159	0.204
pSp	1.57	86.4	0.097
ct2cxr	1.76	**82.3**	**0.075**

Table 1 shows that ct2cxr demonstrates significant improvements over DeepDRR and pSp on the three image quality metrics: Inception Score (IS) [19], FID [7] and KID [3]. The lower IS on pSp and ct2cxr is likely caused by the reduced image diversity due to regularization in exchange for higher fidelity. A visual inspection of Fig. 5 also shows that, in contrast with DeepDRR, ct2cxr is able to generate high resolution CXRs despite low resolution inputs.

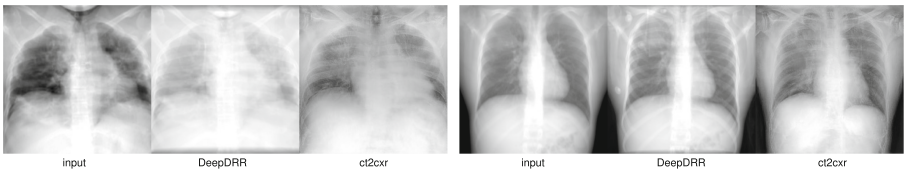


Fig. 5. Comparison between images generated by DeepDRR and ct2cxr. First example shows an input CT with large slice thickness. Second example shows an input CT with small slice thickness.

3.3 Synthesis and Classification Results

A main advantage of the ct2cxr framework over DRR-based CT-CXR conversion is the possibility of style mixing, making it possible to enlarge existing datasets by mixing different input CTs. We created 3 augmented Covid-19 pneumonia CXR datasets (Fig. 6). *Generated* contains 642 CXR images generated through 1–1 CT-CXR synthesis. *Mixed mean* and *mixed 7:7* consist of 2000 CXR images each generated through mean and 7:7 style mixing of random CT pairs.

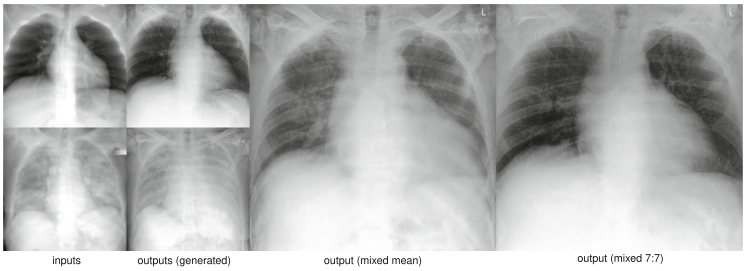


Fig. 6. Images produced by style mixing. *Generated* takes the direct output from the ct2cxr model. *Mean* averages the style vectors, thus equally incorporating features from both images. *7:7* takes the first 7 style vectors from image 1 and the next 7 from image 2, thus preserving large-scale features from image 1.

To test the effectiveness of ct2cxr as a data augmentation tool, we used the images generated by the ct2cxr model to perform further training on a DenseNet121 and ResNet50 pneumonia classification model respectively. The baseline models were trained using our data from a hospital composing 1134 pneumonia positive and 3065 normal CXRs [22]. The testing sets T1 and T2 contain 189 and 72 pneumonia, and 3525 and 1368 normal CXRs labelled by a few radiologists. T3 contains only Covid-19 positive cases confirmed by PCR test, where the 735 positive and 5328 negative pneumonia CXRs (based on radiology reports) contain not only pneumonia and normal (healthy) cases, but also other diseases. The results are shown in Table 2:

Table 2. Classification results. Values represent the Area Under Curve (AUC).

	DenseNet121			ResNet50		
	T1	T2	T3	T1	T2	T3
None (baseline)	0.9446	0.9365	0.8077	0.8957	0.8932	0.7822
CT projection	0.9559	0.9414	0.8031	0.9229	0.9131	0.7946
DeepDRR	0.9515	0.9391	0.8085	0.9002	0.9048	0.7765
Generated	0.9500	0.9423	0.8028	0.9051	0.9148	0.7924
Mixed mean	**0.9588**	0.9432	**0.8149**	0.9275	**0.9368**	**0.8010**
Mixed 7:7	0.9569	**0.9494**	0.8105	**0.9302**	0.9196	0.7897

In the experiments, we observed that the CT projection, DeepDRR and 1–1 Generated augmented data improved the performances of both classification models on T1 and T2, but slightly reduced the T3 AUC of either the DenseNet121 or ResNet50 model. The best performances through the 3 test sets were achieved using Mixed mean and Mixed 7:7, thus demonstrating the capability of our ct2cxr synthesis framework in generating unseen pneumonia cases through style mixing. Note that T3 is a more challenging dataset, which contains negative pneumonia cases with many other diseases, while T1 and T2 mainly contain normal CXRs as negative cases. In particular, compared with DeepDRR, Mixed mean and Mixed 7:7 have t-test p-values of 0.02 and 0.025 respectively, which imply statistically significant performance improvements.

We also implemented the code of [31] and retrained lung swapping autoencoder (LSAE) with our dataset to learn the factorized representation of chest x-ray images. Then, we finetuned the texture encoder in LSAE for further pneumonia classification. The AUC obtained on T1, T2, and T3 datasets are 0.8688, 0.8711, and 0.7626, respectively. Compared with [31], our proposed method has achieved consistently better performances. The lower performance of the model [31] may be due to a few factors of the error introduced by the lung segmentation for severe pneumonia cases and shallow networks of the encoder.

4 Conclusion

In this paper, we have proposed and analyzed a novel framework for CT-based CXR augmentation. We demonstrated that we can improve the performance of an existing pneumonia classification model using the augmented data. Relating to fighting the current pandemic, our framework can be applied to improve the generalizability of existing Covid-19 pneumonia detection models. Furthermore, we originated the idea of using $L2$ loss with the extracted pathology projection as ground truth and using the regularization loss to ensure image fidelity. This concept is applicable to any other pathologies and would enable research on CXR diagnosis models on traditionally CT-diagnosed diseases through effective CT-CXR data augmentation. Further studies would be performed on improving

the capabilities of the pretrained StyleGAN to better synthesize uncommon or small-scale pathologies.

Acknowledgement. The work is partially funded by a GAP project ACCL/19-GAP012-R20H.

References

1. Apostolopoulos, I.D., Mpesiana, T.A.: COVID-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Phys. Eng. Sci. Med. **43**(2), 635–640 (2020)
2. Armato, S.G., III., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd GANs. arXiv preprint arXiv:1801.01401 (2018)
4. Borghesi, A., Maroldi, R.: COVID-19 outbreak in Italy: experimental chest x-ray scoring system for quantifying and monitoring disease progression. Radiol. Med. (Torino) **125**, 509–513 (2020)
5. Chowdhury, M.E., et al.: Can AI help in screening viral and COVID-19 pneumonia? IEEE Access **8**, 132665–132676 (2020)
6. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27**, 1–8 (2014)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Adv. Neural Inf. Process. Syst. **30**, 1–9 (2017)
8. Irvin, J., Rajpurkar, P., Ko, M., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)
9. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)
10. Khan, A.I., Shah, J.L., Bhat, M.M.: CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. Comput. Methods Programs Biomed. **196**, 105581 (2020)
11. Li, X., Yang, J., Zhu, Y.: Digitally reconstructed radiograph generation by an adaptive monte Carlo method. Phys. Med. Biol. **51**(11), 2745 (2006)
12. Liu, S., Dowling, J.A., Engstrom, C., Greer, P.B., Crozier, S., Chandra, S.S.: Manipulating medical image translation with manifold disentanglement. arXiv preprint arXiv:2011.13615 (2020)
13. Loey, M., Smarandache, F., M Khalifa, N.E.: Within the lack of chest COVID-19 x-ray dataset: a novel detection model based on GAN and deep transfer learning. Symmetry. **12**(4), 651 (2020)
14. Maghdid, H.S., Asaad, A.T., Ghafoor, K.Z., Sadiq, A.S., Mirjalili, S., Khan, M.K.: Diagnosing COVID-19 pneumonia from x-ray and CT images using deep learning and transfer learning algorithms. In: Multimodal Image Exploitation and Learning 2021, vol. 11734, p. 117340E. International Society for Optics and Photonics (2021)
15. Mostafavi, S.M.: COVID19-CT-Dataset: An Open-Access Chest CT Image Repository of 1000+ Patients with Confirmed COVID-19 Diagnosis (2021). https://doi.org/10.7910/DVN/6ACUZJ

16. Nguyen, D., et al.: Deep learning-based COVID-19 pneumonia classification using chest CT images: model generalizability. Front. Artif. Intell. **4**, 694875 (2021)
17. Ren, Z., Yu, S.X., Whitney, D.: Controllable medical image generation via generative adversarial networks. Electron. Imaging **2021**(11), 112–121 (2021)
18. Richardson, E., et al.: Encoding in style: a StyleGAN encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2287–2296 (2021)
19. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. Adv. Neural Inf. Process. Syst. **29**, 1–9 (2016)
20. Sherouse, G.W., Novins, K., Chaney, E.L.: Computation of digitally reconstructed radiographs for use in radiotherapy treatment design. Int. J. Rad. Oncol. * Biol.* Phys. **18**(3), 651–658 (1990)
21. Signoroni, A., et al.: BS-Net: learning COVID-19 pneumonia severity on a large chest x-ray dataset. Medical Image Analysis p. 102046 (2021). https://doi.org/10.1016/j.media.2021.102046, https://www.sciencedirect.com/science/article/pii/S136184152100092X
22. Sim, J.Z.T., Ting, Y.H., Tang, Y., et al.: Diagnostic performance of a deep learning model deployed at a national COVID-19 screening facility for detection of pneumonia on frontal chest radiographs. In: Healthcare, vol. 10, p. 175. MDPI (2022)
23. Staub, D., Murphy, M.J.: A digitally reconstructed radiograph algorithm calculated from first principles. Med. Phys. **40**(1), 011902 (2013)
24. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for StyleGAN image manipulation. ACM Trans. Graph. (TOG) **40**(4), 1–14 (2021)
25. Unberath, M., et al.: DeepDRR – a catalyst for machine learning in fluoroscopy-guided procedures. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 98–106. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_12
26. Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., Pinheiro, P.R.: COVIDGAN: data augmentation using auxiliary classifier GAN for improved COVID-19 detection. IEEE Access **8**, 91916–91923 (2020)
27. Wang, L., Chen, W., Yang, W., Bi, F., Yu, F.R.: A state-of-the-art review on image synthesis with generative adversarial networks. IEEE Access **8**, 63514–63537 (2020)
28. Yan, T., Wong, P.K., Ren, H., Wang, H., Wang, J., Li, Y.: Automatic distinction between COVID-19 and common pneumonia using multi-scale convolutional neural network on chest ct scans. Chaos, Solitons Fractals **140**, 110153 (2020)
29. Yousefzadeh, M., et al.: AI-corona: radiologist-assistant deep learning framework for COVID-19 diagnosis in chest CT scans. PLOS ONE. **16**(5), e0250952 (2021)
30. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
31. Zhou, L., et al.: Lung swapping autoencoder: Learning a disentangled structure-texture representation of chest radiographs. arXiv preprint arXiv:2201.07344 (2022)
32. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)



Harmonization of Multi-site Cortical Data Across the Human Lifespan

Sahar Ahmad[1], Fang Nan[2], Ye Wu[1], Zhengwang Wu[1], Weili Lin[1], Li Wang[1], Gang Li[1], Di Wu[3,4], and Pew-Thian Yap[1(✉)]

[1] Department of Radiology and Biomedical Research Imaging Center (BRIC), The University of North Carolina at Chapel Hill, NC, USA
ptyap@med.unc.edu
[2] Department of Biostatistics, University of Washington, Seattle, WA, USA
[3] Department of Biostatistics, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, NC, USA
[4] Division of Oral and Craniofacial Health Research, Adams School of Dentistry, The University of North Carolina at Chapel Hill, NC, USA

Abstract. Neuroimaging data harmonization has become a prerequisite in integrative data analytics for standardizing a wide variety of data collected from multiple studies and enabling interdisciplinary research. The lack of standardized image acquisition and computational procedures introduces non-biological variability and inconsistency in multi-site data, complicating downstream statistical analyses. Here, we propose a novel statistical technique to retrospectively harmonize multi-site cortical data collected longitudinally and cross-sectionally between birth and 100 years. We demonstrate that our method can effectively eliminate non-biological disparities from cortical thickness and myelination measurements, while preserving biological variation across the entire lifespan. Our harmonization method will foster large-scale population studies by providing comparable data required for investigating developmental and aging processes.

Keywords: Harmonization · Lifespan · Multi-site · Cortex

1 Introduction

Data pooling and sharing is becoming common in neuroimaging studies [14,18], leveraging diversified data to elucidate the developmental and aging processes throughout the human lifespan [6]. Open data archiving also promotes research transparency, reproducibility, and rigor. Given the large-scale multi-site neuroimaging data, integrative data analytics affords greater statistical power in analyzing the dynamic evolutionary changes of multiple brain characteristics.

Although neuroimaging big data are key to deciphering the brain architecture at multiple scales spatiotemporally, integrating and mining the massive multi-site data is challenging due to non-biological differences resulting from the use

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 220–229, 2022.
https://doi.org/10.1007/978-3-031-21014-3_23

of (i) study-specific image acquisition protocols; (ii) different scanner hardware and software; and (iii) non-standardized post-acquisition image preprocessing techniques. These scanner-induced variabilities and inconsistent computational procedures confound biological effects in the collected data; thus, introducing inconsistencies in downstream statistical analyses.

To address these issues, data harmonization is employed to model non-biological variabilities and eliminate the corresponding bias. ComBat, originally developed to harmonize genomic data [12], was first applied to harmonize diffusion tensor imaging measurements [5]. It was subsequently modified to accommodate diverse neuroimaging data, including brain regional volumes [17], cortical thickness [2,4], and radiomic features [16]. While the variants of ComBat are effective in removing non-biological variations, they are not necessarily effective in dealing with longitudinal or cross-sectional data covering the entire human lifespan.

In this paper, we propose a statistical harmonization method that accounts for site-specific non-biological variability in both longitudinal and cross-sectional data collected between birth and centenary. We integrated cortical thickness and myelination measurements from five different connectome studies, covering the entire human lifespan. We show that our method can effectively and reliably remove unwanted site-related effects while at the same time capture nonlinear age-related variations across the lifespan.

2 Materials and Methods

2.1 Datasets and Preprocessing

Structural MRI data consisting of T1-weighted (T1w), T2-weighted (T2w), and white and pial cortical surfaces were pooled from five connectome studies: Developing Human Connectome Project (dHCP) [15]; Baby Connectome Project (BCP) [9]; HCP Development (HCP-D); HCP Young Adult (HCP-YA); and HCP Aging (HCP-A) [7,19]. Our lifespan dataset comprises of both longitudinally and cross-sectionally collected samples; the age range and number of scans for each study are shown in Fig. 1a. We used minimally preprocessed data from dHCP, HCP-D, HCP-YA, and HCP-A [3,10,11] and BCP data were preprocessed via iBEAT[1] [13,20].

We measured cortical thickness between white and pial surfaces (Fig. 1b), and intracortical myelination as a ratio between T1w and T2w intensities within the cortical ribbon of each scan (Fig. 1c). The computed T1w/T2w ratio was then mapped onto individual white surface using Connectome Workbench[2] [8].

2.2 Harmonization of Cortical Data

We propose to harmonize cortical thickness and cortical myelination across the entire lifespan via our new statistical method called piecewise adjustment

[1] https://ibeat.wildapricot.org/.
[2] https://www.humanconnectome.org/software/connectome-workbench.

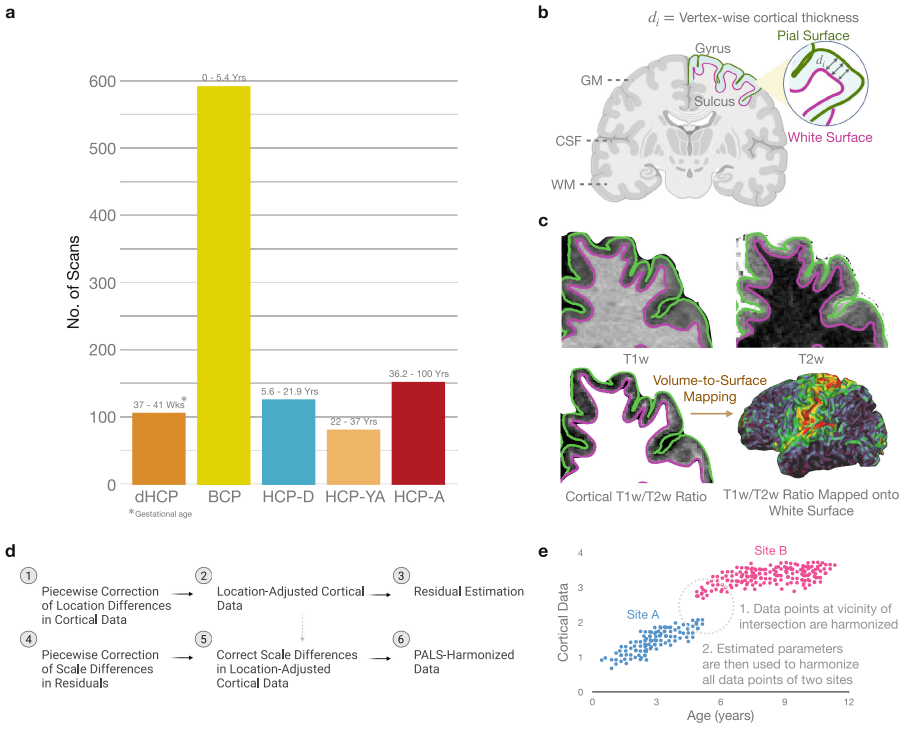


Fig. 1. (a) Frequency distribution of data pooled from different neuroimaging studies. (b) Illustration of cortical thickness measured between white and pial cortical surfaces. (c) Intracortical myelination mapped from volume to white surface of select sample. (d) PALS harmonization steps. (e) Piecewise adjustment strategy employed in PALS.

of location and scale (PALS), which is partially based on ComBat [12]. Our method adjusts for location (mean) and scale (variance) differences in multi-site cortical data, while preserving biological variations in the data. To accommodate dynamic age-related variations in cortical data, we incorporate a piecewise adjustment approach in PALS, which at each time processes data points in close proximity to the intersection of two sites (Fig. 1e). PALS is a two-step approach: (i) site-related location differences are corrected from raw cortical data; and (ii) scale differences are corrected from residuals of the adjusted cortical data obtained in previous step. We illustrate the steps involved in PALS in Fig. 1d and detail them below.

Adjustment of Location Effects. Let $Y_{ijk}(t)$ be the cortical measurement for scan number t[3], site i, subject j, and surface vertex k. We define the generalized additive mixture model (GAMM) $f_k(X_{ij})$ that captures age-related (X)

[3] $t = 1$ for cross-sectional data; $t \geq 1$ for longitudinal data

variations in Y as:

$$Y_{ijk}(t) = f_k(X_{ij}) = f(X_{ij}) + \gamma_{jk}, \tag{1}$$

where $f(X_{ij})$ is a smooth nonlinear function of age and γ_{jk} is the subject-specific random intercept. We employ the empirical Bayes (EB) framework [12] as proposed in ComBat to correct location differences (g_{ik}) in the cortical data. The magnitude of cortical data differs across surface vertices and if these differences are not accounted for, the prior estimates of g_{ik} in EB framework will be biased. To overcome this, we first fit GAMM to cortical data to estimate overall mean effect and then standardize the vertex-specific data, such that the mean and variance of cortical data for all vertices is similar. The standardized data $Z_{ijk}(t)$ are given by

$$Z_{ijk}(t) = \frac{Y_{ijk}(t) - f_k(X_{ij})}{s_k^2}, \tag{2}$$

where s_k^2 is the vertex-specific variance. Assuming normal distribution as prior for location-specific differences $g_{ik} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, EB framework empirically estimates the hyperparameters of the prior distribution from the standardized data. Using these hyperparameters, the conditional posterior estimates of location effects (g_{ik}^*) are computed and applied to obtain the location-corrected cortical data ($\hat{Y}_{ijk}(t)$):

$$\hat{Y}_{ijk}(t) = [Z_{ijk}(t) - g_{ik}^*]s_k^2 + f_k(X_{ij}). \tag{3}$$

Adjustment of Scale Effects. The location-corrected cortical data $\hat{Y}_{ijk}(t)$ are adjusted for site-specific scale differences d_{ik}. To accommodate age effect on the variance, PALS harmonizes data in two phases: (i) scale differences are estimated from absolute residuals of $\hat{Y}_{ijk}(t)$; and (ii) estimated scale differences are applied to $\hat{Y}_{ijk}(t)$ to obtain PALS-harmonized data $Y_{ijk}^*(t)$. Let $R_{ijk}(t)$ be the absolute residual of $\hat{Y}_{ijk}(t)$ obtained by fitting GAMM. Assuming inverse gamma distribution as prior for scale differences $d_{ik} \sim \mathcal{IG}(\lambda_i, \theta_i)$, EB framework is employed to estimate the hyperparameters of the prior distribution. These hyperparameters are used to compute the conditional posterior estimates of scale effects (d_{ik}^*) that give $Y_{ijk}^*(t)$ as

$$Y_{ijk}^*(t) = \left[\frac{\hat{Z}_{ijk}(t)}{d_{ik}^*}\right] \hat{s}_k^2 + \hat{f}_k(X_{ij}), \tag{4}$$

where $\hat{Z}_{ijk}(t)$ are the standardized data given by

$$\hat{Z}_{ijk}(t) = \frac{\hat{Y}_{ijk}(t) - \hat{f}_k(X_{ij})}{\hat{s}_k^2}. \tag{5}$$

2.3 Growth Modeling of Cortical Data

We use PALS-harmonized cortical data to model the development and maturation patterns across the lifespan. For this, we first use sliding window technique

to correct outliers in the harmonized data of each scan as follows:

$$\mathcal{Y}_{ijl}(t) = \begin{cases} \text{median}(W^*_{ij}(t)) + a \times \text{mad}(W^*_{ij}(t)); & \text{if } W^*_{ijl}(t) > T_1, \\ \text{median}(W^*_{ij}(t)) - a \times \text{mad}(W^*_{ij}(t)); & \text{if } W^*_{ijl}(t) < T_2, \end{cases} \quad (6)$$

where $a = 1$ and $W^*_{ij}(t)$ are the harmonized cortical data within the current window of width = 300 vertices; l is the vertex index within the sliding window; median$(W^*_{ij}(t))$ and mad$(W^*_{ij}(t))$ are the median and median absolute deviation computed over $W^*_{ij}(t)$ for the j-th subject and site i, respectively. The thresholds are defined as $T_1 = \text{median}(W^*_{ij}(t)) + c \times \text{mad}(W^*_{ij}(t))$ and $T_2 = \text{median}(W^*_{ij}(t)) - c \times \text{mad}(W^*_{ij}(t))$, where $c = 2$. Next, we model the outlier corrected data $(\mathcal{Y}_{ij}(t))$ using GAMM with cubic regression spline as smooth nonlinear function of age and subject-specific random intercept. The degree of smoothness is automatically selected with the restricted maximum likelihood (REML) criterion [1] in R software.

3 Results and Discussion

We compared our PALS-harmonized cortical data with raw data and harmonized data obtained via (i) longitudinal ComBat (longCombat) [2] and (ii) ComBat-GAM [17]. We present the mean cortical thickness and cortical myelination before and after harmonization in Fig. 2. PALS effectively removed site-related location and scale effects from both the hemispheric cortical data. The residuals given by PALS (Fig. 3) are centered around zero and show a gradual scale change with age for both cortical measurements. These results demonstrate that longCombat and ComBat-GAM did not completely eliminate non-biological confounds, causing the data to disperse. This is because they are not designed to accommodate both longitudinal and cross-sectional data demonstrating dynamic age changes across the entire lifespan.

To check the efficacy of harmonization methods in preserving biological variability, we quantified proportion of variation captured by age in raw and harmonized data. For this, we performed regression using a hybrid model with median cortical measurement as response variable $(M = (1 - \text{e}^{(-\text{age}/\beta_0)}) + \beta_1 \text{age} + \beta_2 \text{age}^2 + \beta_3 \text{age}^3)$ and computed coefficient of variation R^2. The results in Table 1 show that the percentage of variation explained by age is highest for PALS for both cortical measurements, validating that PALS successfully removed confounding site effects while preserving biological variability.

Next, we fitted GAMM to PALS-harmonized data to obtain growth trajectories. We visualize the lifespan developmental and maturational pattern for cortical thickness and myelination on age-specific white surfaces of the inflated atlases (Fig. 4). We observe an inverse relationship between thickness and myelination, in line with [8]: thinner cortices are heavily myelinated compared to thicker cortices. For instance, primary sensorimotor and visual cortices being thinner are more myelinated than the rest of the brain. Thus, PALS also maintained neurobiological relationship between the two cortical features.

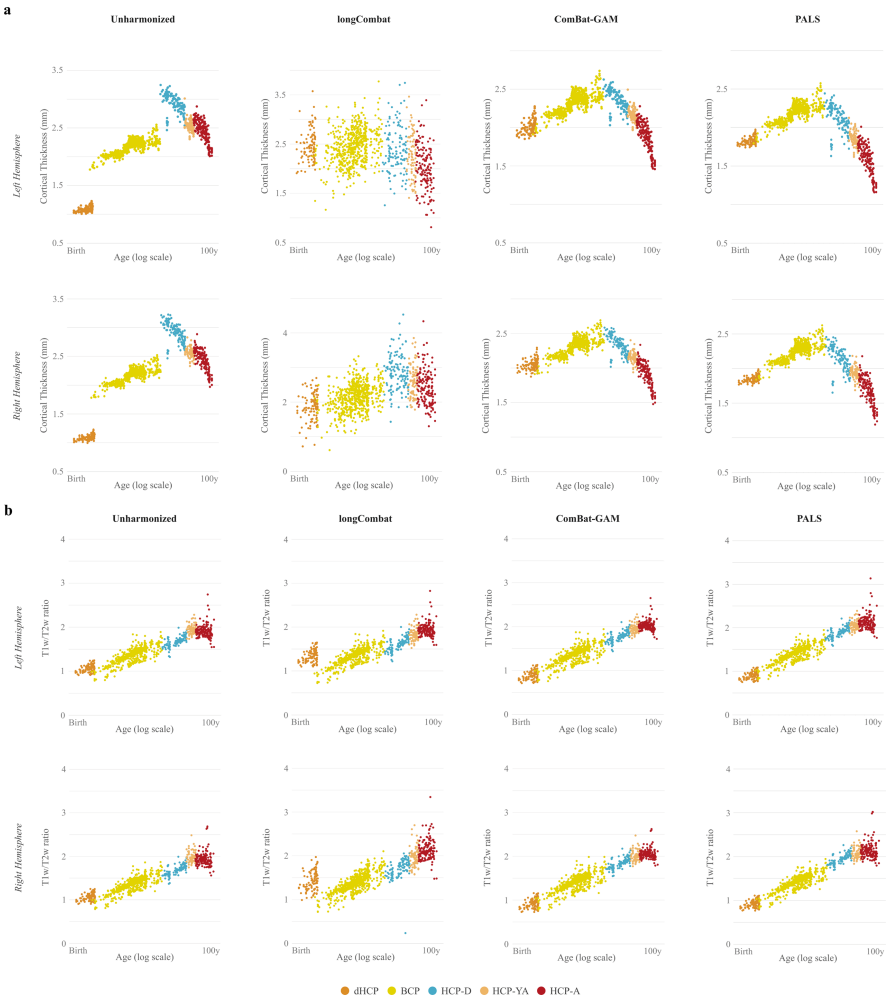


Fig. 2. Comparison of (a) cortical thickness and (b) cortical myelination before and after harmonization.

Table 1. Coefficient of variation R^2 obtained before and after harmonization.

Cortical data	Unharmonized	longCombat	ComBat-GAM	PALS
Left cortical thickness	0.80	0.76	0.85	0.88
Right cortical thickness	0.80	0.69	0.77	0.86
Left cortical myelination	0.87	0.75	0.90	0.92
Right cortical myelination	0.88	0.76	0.90	0.91

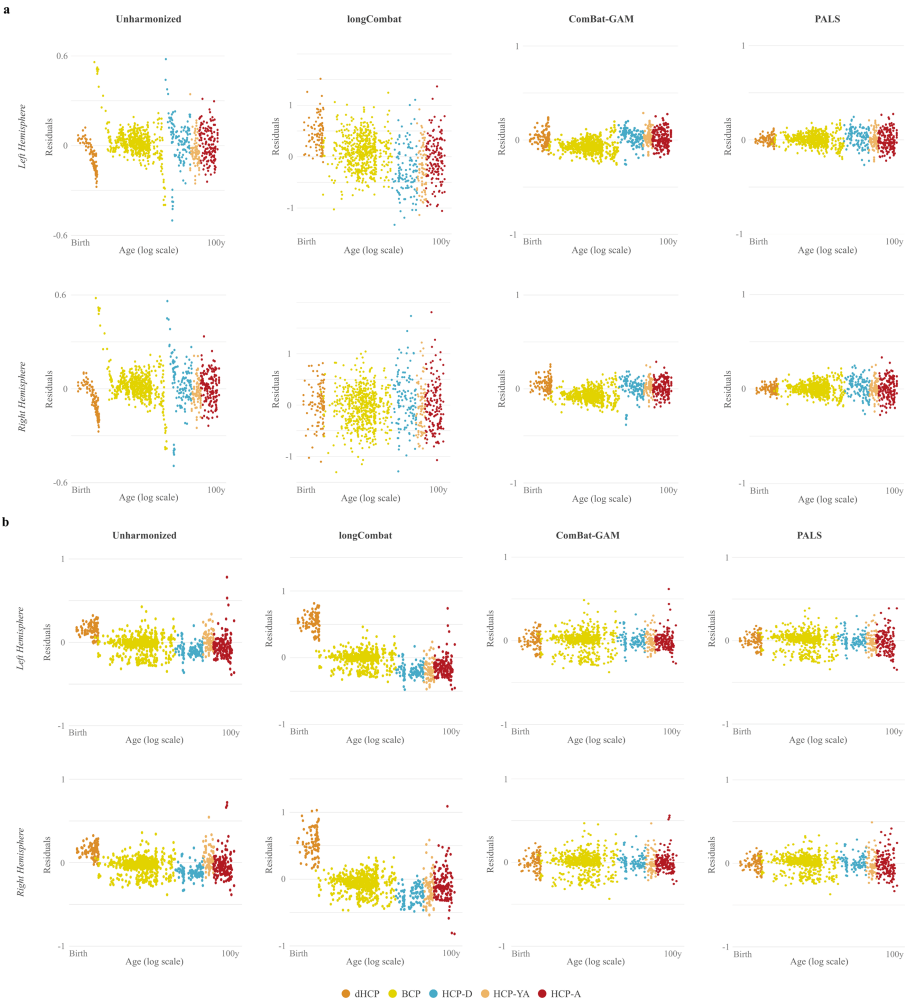


Fig. 3. Comparison of residuals for (a) cortical thickness and (b) cortical myelination before and after harmonization.

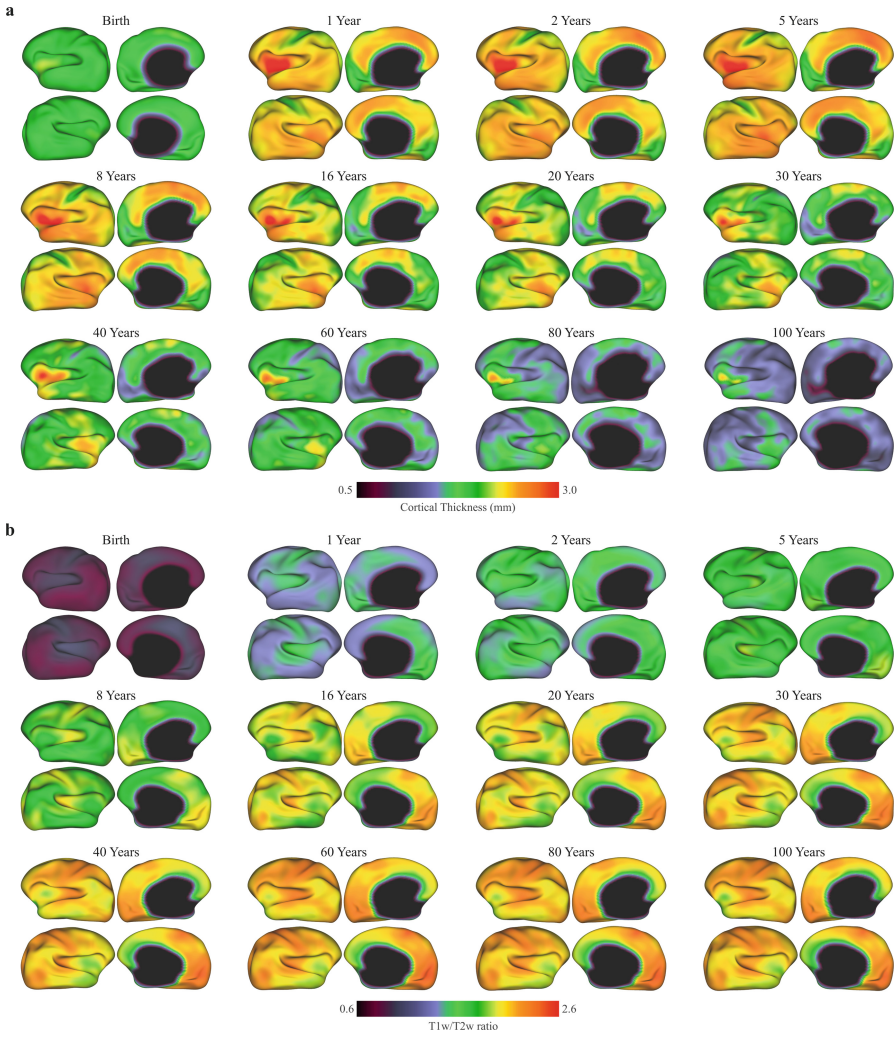


Fig. 4. PALS-harmonized (a) cortical thickness and (b) cortical myelination mapped onto inflated age-specific atlases of the white surface.

4 Conclusion

Our proposed harmonization method, PALS in conjunction with GAMM, outperformed existing harmonization methods in removing site-specific non-biological discrepancies and accurately modeled age-related dynamic variability in lifespan longitudinal and cross-sectional cortical data. Thus, our method can be employed in large-scale population-based studies to explore brain structure and function.

Acknowledgments. This work was supported in part by the United States National Institutes of Health (NIH) through grants MH125479 and EB008374.

References

1. Bartlett, M.S.: Properties of sufficiency and statistical tests. Proc. R. Soc. London. Ser. A Math. Phys. Sci. **160**(901), 268–282 (1937). www.jstor.org/stable/96803
2. Beer, J.C., et al.: Longitudinal ComBat: a method for harmonizing longitudinal multi-scanner imaging data. Neuroimage **220**, 117129 (2020). https://doi.org/10.1016/j.neuroimage.2020.117129
3. Fischl, B.: Freesurfer. Neuroimage **62**(2), 774–781 (2012). https://doi.org/10.1016/j.neuroimage.2012.01.021
4. Fortin, J.P., et al.: Harmonization of cortical thickness measurements across scanners and sites. Neuroimage **167**, 104–120 (2018). https://doi.org/10.1016/j.neuroimage.2017.11.024
5. Fortin, J.P., et al.: Harmonization of multi-site diffusion tensor imaging data. Neuroimage **161**, 149–170 (2017). https://doi.org/10.1016/j.neuroimage.2017.08.047
6. Frangou, S., et al.: Cortical thickness across the lifespan: data from 17,075 healthy individuals aged 3–90 years. Hum. Brain Mapp. **43**(1), 431–451 (2022). https://doi.org/10.1002/hbm.25364
7. Glasser, M.F., et al.: The minimal preprocessing pipelines for the human connectome project. Neuroimage **80**, 105–124 (2013). https://doi.org/10.1016/j.neuroimage.2013.04.127
8. Glasser, M.F., Van Essen, D.C.: Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. J. Neurosci. **31**(32), 11597–11616 (2011). www.jneurosci.org/content/31/32/11597
9. Howell, B.R.,et al.: The UNC/UMN baby connectome project (BCP): an overview of the study design and protocol development. Neuroimage. **185**, 891–905 (2019). https://doi.org/10.1016/j.neuroimage.2018.03.049
10. Jenkinson, M., Bannister, P., Brady, M., Smith, S.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage **17**(2), 825–841 (2002). https://doi.org/10.1006/nimg.2002.1132
11. Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M.: FSL. Neuroimage **62**(2), 782–790 (2012). https://doi.org/10.1016/j.neuroimage.2011.09.015
12. Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics **8**(1), 118–127 (2006). https://doi.org/10.1093/biostatistics/kxj037
13. Li, G., et al.: Measuring the dynamic longitudinal cortex development in infants by reconstruction of temporally consistent cortical surfaces. Neuroimage **90**, 266–279 (2014). https://doi.org/10.1016/j.neuroimage.2013.12.038
14. Littlejohns, T.J., et al.: The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat. Commun. **11**, 2624 (2020). https://doi.org/10.1038/s41467-020-15948-9
15. Makropoulos, A., et al.: The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. Neuroimage **173**, 88–112 (2018). https://doi.org/10.1016/j.neuroimage.2018.01.054
16. Orlhac, F., et al.: A postreconstruction harmonization method for multicenter radiomic studies in pet. J. Nucl. Med. **59**(8), 1321–1328 (2018). https://doi.org/10.2967/jnumed.117.199935

17. Pomponio, R., et al.: Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage **208**, 116450 (2020). https://doi.org/10.1016/j.neuroimage.2019.116450
18. Thompson, P.M., Stein, J.L., Medland, S.E., Hibar, D.P., et al.: The ENIGMA consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. **8**, 153–182 (2014). https://doi.org/10.1007/s11682-013-9269-5
19. Van Essen, D., et al.: The human connectome project: a data acquisition perspective. Neuroimage **62**(4), 2222–2231 (2012). https://doi.org/10.1016/j.neuroimage.2012.02.018
20. Wang, L., et al.: Volume-based analysis of 6-month-old infant brain MRI for Autism biomarker identification and early diagnosis. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11072, pp. 411–419. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_47



Head and Neck Vessel Segmentation with Connective Topology Using Affinity Graph

Linlin Yao[1,2], Zhong Xue[1], Yiqiang Zhan[1], Lizhou Chen[4], Yuntian Chen[4], Bin Song[4], Qian Wang[3(✉)], Feng Shi[1(✉)], and Dinggang Shen[1,3]

[1] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
feng.shi@uii-ai.com
[2] Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[3] School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
wangqian2@shanghaitech.edu.cn
[4] Department of Radiology, West China Hospital, Sichuan University, Chengdu, China

Abstract. Vessel segmentation is an important step for cerebrovascular disease analysis, while automatic and complete segmentation of head and neck vessels in CT angiography is a challenging problem. The reason is that vessels have diverse shapes and sizes in long and tortuous tubular-like vasculatures, as well as confounding appearance with surrounding tissues. Current deep learning-based methods often use voxel-wise segmentation, without considering the shape or connectivity of segmented vessels. In this paper, we describe vascular structures using affinity maps and construct connectivity priors of vessels so as to establish a topological connectivity loss for vessel segmentation. Specifically, a multi-head 3D U-Net is applied to predict the segmentation mask and the affinity map, and subsequently the connectivity-aware affinity map is used to refine the segmentation. In experiments, we applied our method on 72 head and neck CT angiography images. Voxel-wise and topology-relevant metrics show that the proposed method achieved superior performance than the widely used 3D nnU-Net and provided better 3D vessel visualization results.

Keywords: Vessel topological connectivity · Head and neck vessel segmentation · Affinity map · CT Angiography

1 Introduction

Cerebrovascular disease is one of the leading causes of severe disability and mortality worldwide [1]. Computer Tomography Angiography (CTA) is a widely used, minimally invasive, and cost-efficient imaging modality in routine clinical diagnoses of vascular diseases [2]. To visually inspect and quantitatively analyze the vasculature of head and neck more efficiently, vessel is usually segmented and constructed to 3D visualization for clinical examination. However, manual

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 230–238, 2022.
https://doi.org/10.1007/978-3-031-21014-3_24

delineation of head and neck vessels is time-consuming and labor-intensive even for experienced radiologists, as they are long and tortuous tubular-like vascular structure with diverse shapes and sizes. Thus, automatic and accurate segmentation for head and neck vessels is highly desirable to simplify and speed up the clinical 3D vessel reconstruction process.

Deep learning-based segmentation approaches have drawn increasing interests due to their self-learning and generalization abilities from large number of data volumes. Currently, there are a few works [3–5] about head and neck vessels segmentation from CTA images. Briefly, deep networks were developed in these studies to predict dense probability maps indicating the percentage of each voxel belongs to vessels, and then the map is thresholded to obtain the segmentation. Metrics, such as Dice similarity Coefficient (DSC) are used for evaluation. However, these metrics cannot quantify the tubular topology or vessel connectivity of segmentation. The resultant vessels are often suffered from structural discontinuity.

To improve connectivity of vessel segmentation, Xu et al. [6] proposed a recursive semantics-guided network for better connectivity on retinal vessel segmentation. Paetzold et al. [7] proposed a connectivity-preserving metric clDice for vessel segmentation, where centerline dice is computed voxel-wisely. On the other hand, structural continuity methods are investigated. Funke et al. [8] presented a method combining affinity prediction with region agglomeration for connectome reconstruction on electron microscopy volumes. Hu et al. [9] proposed a topology-preserving loss function for segmenting retinal image with correct topology. Qin et al. [10] presented a voxel-connectivity aware approach named AirwayNet for airway segmentation, in which binary segmentation task is transformed into 26 tasks of connectivity prediction.

Inspired by these works, we propose a novel method for predicting a segmentation mask and improving the structural continuity simultaneously by using affinity graph. The advantage is that vessel segment connectivity information can be considered as an affinity connection problem. This method not only improves the voxel-wise segmentation performance, but also obtains result with better topological connectivity. We evaluate our proposed method by both voxel-wise and topological connectivity metrics.

2 Method

Figure 1 gives an overview of the proposed method. There are three main steps: 1) Affinity graph construction from ground truth mask; 2) Network training for learning the segmentation mask and the affinity map; 3) Affinity-based connectivity-aware segmentation refinement. Each step is described in the following subsections in detail.

2.1 Affinity Graph Construction

Affinity graph ground truth A used as supervision in training process is constructed from the vessel mask Y, with size of $L \times W \times H$, shown in Fig. 1. Based

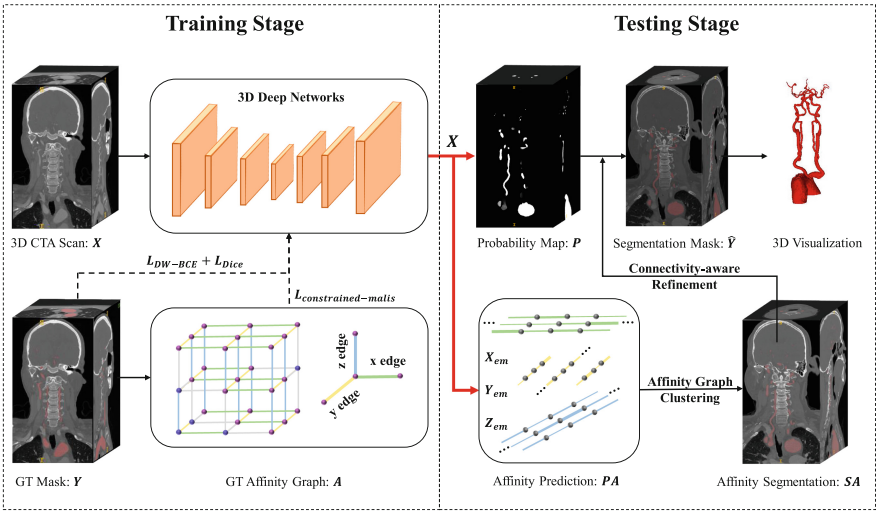


Fig. 1. The training and testing stages of the method. It has two head branches. The first one is targeting on the ground-truth label map, which performs per-voxel vessel/non-vessel classification. The second head branch is targeting on the affinity map estimation. Then, to leverage the connectivity prior obtained from the affinity graph, an clustering approach is proposed to generate an over-segmentation and followed by connectivity-aware refinement.

on the coordinate set $V = \{v | v \in L \times W \times H\}$, an affinity graph $G = (V, E, a)$ with nodes V and edges E weighted by affinities a can be defined. In three dimensional image, affinity maps of three directions (right, anterior, superior directions corresponding to x, y, z axes) can be obtained according to the definition: for every pair of voxels there exist an edge if and only if they are neighbors. An affinity of 1 between pairs of neighboring nodes is set when they have the same label in the vessel mask Y. On the opposite, an affinity of 0 means they belong to the different regions. The graph G which is six-connected can be constructed from the three directional affinity maps. As shown in the ground-truth affinity graph A in Fig. 1, purple and pink nodes with different labels represent different regions. Affinities of 1 in affinity maps are represented by green (x direction), yellow (y direction) and blue (z direction) edges respectively and affinities of 0 are described by gray edges.

2.2 Network Training for Segmentation

Given a 3D CTA scan X and its ground-truth label map Y, we can compute its affinity graph A according to the method given in the above section. In this section, we train a deep network for head and neck vessel segmentation by targeting on both Y, using distance weighted binary cross-entropy(DW-BCE) named as $\mathcal{L}_{cls}$ and dice loss $\mathcal{L}_{dice}$, and A by calculating the structured topological loss on the affinity graph.

Semantic Segmentation. As shown in Fig. 1, our model has two head branches. The first one is targeting on the ground-truth mask Y, which performs per-voxel classification for semantic segmentation with the distance weighted binary cross-entropy (DW-BCE) loss function $\mathcal{L}_{cls}$ and dice loss $\mathcal{L}_{dice}$.

$$\mathcal{L}_{cls} = -\sum_{v \in V}(\beta_p y_v log p_v(W, w_{cls}) + \beta_n(1 - y_v) log(1 - p_v(W, w_{cls}))), \quad (1)$$

where W is the parameters of the network backbone, w_{cls} is the parameters of this head branch and $p_v(W, w_{cls})$ is the probability that v is a vessel voxel. $\beta_p = \frac{D_v}{\sum y_v}$ and $\beta_n = \frac{D_v}{\sum (1-y_v)}$. D_v represents the distance weight of voxel v. Distance weighted map was generated by computing the distance transform on the segmentation masks. The map was then reverted to have positive value inside the segmentation and then inversed to create a distance mask where voxels in the surface of the tubular structure have higher weight, compared to those located in the center.

Affinity Map Estimation. The output of the second head branch consists of three directional affinity maps, which can construct the affinity graph A. The loss function used to update parameters of this branch is adopted from the constrained malis loss [8] defined as following:

$$\mathcal{L}_{aff} = \sum_{e \in MST(Y)} w_P(e) l(1, a_v(e, W, w_a)) + w_N(e) l(0, a_v(e, W, w_a)), \quad (2)$$

where W represents the parameter of the backbone, w_a refers to the parameter of the second head branch, e represents edge between nodes, a is the edge's affinity weight.

Instead of calculating the loss over all the voxel pairs, positive and negative passes are defined base on [8], and w_P and w_N are the weighting parameters of them. Given the ground truth, the positive predicted affinity map is modified by setting the original predicted affinity values of edges crossing different ground truth regions to zero. On the other hand, the negative predicted affinity map is obtained by setting the original affinities within each ground truth region to one. Then, the maximal spanning trees (MST) are constructed by identifying the maximin edges from the positive and the negative predicted affinity maps based on Kruskal's algorithm [11], respectively, resulting in a positive pass and a negative pass.

Affinity of each maximin edge between two voxels m and n is defined based on the maximin path $P^*_{m,n}$, *i.e.,*

$$P^*_{m,n} = \arg \max_{p \in \mathcal{P}_{m,n}} \min_{e \in p} a(e), \quad (3)$$

where $\mathcal{P}_{m,n}$ denotes the set of all possible paths connecting m and n, e is an edge along path p ($p \in \mathcal{P}_{m,n}$), and a is the affinity of edge e.

Consequently, the constrained malis loss is formed by the positive and the negative passes according to their affinities of maximin edges on the MSTs,

wherein the positive pass contributes the loss within regions, and the negative pass focuses on disconnected areas crossing regions.

In summary, the loss function for our segmentation network is:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{dice} + \mathcal{L}_{aff}. \tag{4}$$

2.3 Connectivity-Aware Refinement

Given a 3D CT scan X in the testing phase, for each voxel v, our vessel segmentation network has two branches. The first branch outputs the probability map P, and the second branch outputs three directions affinities maps shown as affinity prediction PA in Fig. 1, which can form the affinity graph.

Affinity Graph Clustering. Once obtaining the affinity graph, hierarchical agglomeration of segments from the predicted affinities can be conducted to generate the affinity connective segmentation SA. It consists of two steps: First, we use a seeded watershed to extract small fragments directly from the predicted affinities. For the seeded watershed, we first average the predicted affinities for each voxel to obtain a volume of boundary predictions. We subsequently threshold the boundary predictions at 0.5 and perform a distance transform on the resulting mask. Every local maximum is taken as a seed, from which we grow basins using watershed on the boundary predictions. After this step, the resulting segmentation still includes many small regions. Second, we use a fragment agglomeration algorithm to fuse regions together. For the fragment agglomeration, we iteratively score and merge adjacent fragments into larger regions until a predefined threshold is reached. Briefly, we first construct a region adjacency graph. Then we merge regions based on the edge weight (the predicted affinity) between them. If two regions are merged, we update the weights with neighboring regions.

Refinement. After getting the predicted affinity connective segmentation SA and the predicted semantic probability map P, the final segmentation mask $\hat{Y}$ is obtained by changing the voxel label with low probability threshold θ in the semantic segmentation, according to the SA when the six-neighbors of the voxel are in the same connected component.

3 Results

3.1 Materials and Parameter Settings

In our experiment, CTA images of 72 subjects were used to evaluate the proposed method. The size of the CTA image is $512 \times 512 \times 533$ and the voxel spacing is $0.4063 \times 0.4063 \times 0.7000$ mm^3. Images covered both head and neck, and vessels were manually annotated by a senior radiologist.

Our implementation was based on PyTorch. Experiments were performed under five-fold cross-validation. We used 3D nnU-Net as the baseline and randomly sampled patches of a specified size (i.e., 128). We trained the network

for 96 epochs using Adam optimizer with a learning rate of 0.001, the distance weighted binary cross-entropy loss and the dice loss. We used 3D U-Net architectures [12] with four levels of different resolutions. Patches size was set as 64 here. Probability threshold θ was set as 0.15 to revise the semantic segmentation according to the affinity segmentation.

3.2 Evaluation Criteria

We evaluate our method by calculating voxel-wise and topological connectivity metrics. Voxel-wise metrics include Dice similarity coefficient (DSC), 95^{th} percentile hausdorff distance (95HD) and average Hausdorff distance (AVD). For better evaluation, we extracted the vessel centerlines through OpenCV skimage package. Metrics for centerline consist of completeness (Compl.), correctness (Corr.) and quality (Qual.). Topological connectivity metrics contains betti error, adapted rand index (ARI) and variation of Information (VOI). Details are as follows.

The Betti error directly compares the topology (number of handles) between the segmentation and the ground truth. Lower the Betti error means better results.

Adapted Rand Index (ARI) is the maximal F-score of the foreground-restricted Rand index. The Rand index $(1 - RI(\hat{Y}, Y))$ measuring the similarity between two clusters is defined as following:

$$1 - RI(\hat{Y}, Y) = \binom{N}{2}^{-1} \sum_{i<j} l(\delta(y_i, y_j), \delta(\hat{y}_i, \hat{y}_j), \tag{5}$$

where Y and $\hat{Y}$ represent the ground truth and the prediction segmentation result, separately. The indicator function $\delta(y_i, y_j)$ is 1 if pixels i and j belong to the same segment ($y_i = y_j$) and 0 otherwise. Rand index ranges from 0 to 1, where higher is better.

The Variation of Information (VOI) is determined by:

$$VOI(\hat{Y}, Y) = H(\hat{Y}|Y) + H(Y|\hat{Y}), \tag{6}$$

where the conditional entropy $H(\hat{Y}|Y)$ is corresponding to the VOI split, and can be interpreted as the amount of over-segmentation. Likewise, the conditional entropy $H(Y|\hat{Y})$ is corresponding to the VOI merge, as the amount of under-segmentation. The VOI would be lower for better results.

Centerline evaluation metrics are defined as following in our experiment:

$$corr. = \frac{|\ \hat{Y}_{cl} \bigcap Y\ |}{|\ \hat{Y}_{cl}\ |} \quad compl. = \frac{|\ Y_{cl} \bigcap \hat{Y}\ |}{|\ Y_{cl}\ |} \quad qual. = \frac{|\ Y_{cl} \bigcap \hat{Y}\ | + |\ \hat{Y}_{cl} \bigcap Y\ |}{|\ Y_{cl}\ | + |\ \hat{Y}_{cl}\ |}, \tag{7}$$

where compl., corr., and qual. are short for completeness, correctness and quality. We deonte the centerline extracted from the ground-truth Y as Y_{cl}, and centerline from the final predicted mask $\hat{Y}$ as $\hat{Y}_{cl}$.

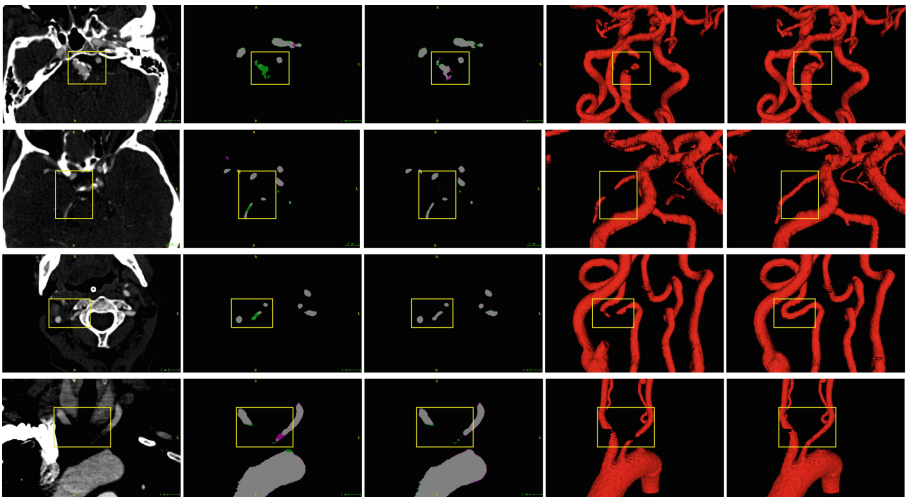


Fig. 2. Qualitative results. From left to right is original CTA, comparison of nnU-Net with ground-truth, comparison of the proposed with ground-truth, nnU-Net segmentation 3D visualization, and the proposed segmentation 3D visualization. In comparison columns (second and third): overlapped-gray; ground-truth-green; automatic-magenta. The yellow rectangles on the figures highlight the comparison regions. (Color figure online)

3.3 Results

Qualitative Results. Qualitative results of 4 subjects are demonstrated in Fig. 2. The corresponding original images are shown in the first column. The segmentation results of nnU-Net [13] are shown in the second and fourth columns, and the results of proposed method are shown in the third and fifth columns. In the second and third comparison columns, gray region means the overlay of ground-truth (green) and automatic segmentation (magenta). 3D visualizations of nnU-Net and the proposed method are demonstrated in fourth and fifth columns, which can provide an intuitive comparison of vessel construction results. By analyzing these results, we can observe that the proposed method performs well in the vessel region influenced by nearby bone tissue and even with low imaging quality. In summary, qualitative results illustrate that the proposed method has better performance on vessel connectivity and more accurate than that of nnU-Net.

Quantitative Results. Quantitative statistic results are presented in Table 1. Both voxel-wise metrics (volume and centerline rows) and topology-relevant metrics (connectivity rows) are calculated to evaluate our proposed method. For volume metrics, compared to baseline, the method improves about 0.008 at DSC, 3.937 at $95HD$ and around 1.71 at AVD. Note that there is one segmentation of nnU-Net having discontinuous vessel segmentation with a number of isolated

Table 1. The quantitative evaluation results.

Volume		DSC	95HD	AVD
	nnU-Net (baseline)	0.973 ± 0.014	5.502 ± 8.197	2.594 ± 1.390
	Proposed	$\mathbf{0.981 \pm 0.008}$	$\mathbf{1.565 \pm 0.427}$	$\mathbf{0.884 \pm 0.388}$
Centerline		Corr	Compl	Qual
	nnU-Net (baseline)	0.913 ± 0.026	0.895 ± 0.051	0.883 ± 0.089
	Proposed	$\mathbf{0.949 \pm 0.017}$	$\mathbf{0.935 \pm 0.017}$	$\mathbf{0.923 \pm 0.043}$
Connectivity		Betti Error	ARI	VOI
	nnU-Net (baseline)	3.187 ± 0.307	0.943 ± 0.0087	1.235 ± 0.025
	Proposed	$\mathbf{1.129 \pm 0.104}$	$\mathbf{0.944 \pm 0.007}$	$\mathbf{0.782 \pm 0.019}$

pieces, which affects the average $95HD$. For centerline metrics that evaluate the tubular structure, correctness, completeness and quality are improved by 0.036, 0.04, and 0.04 respectively. It can be seen that the proposed method has better performance at volume and structure levels. Considering the topology-relevant metrics, Betti Error, ARI and VOI are improved by 2.058, 0.001, and 0.453 respectively. Results show that the proposed method has better vessel segmentation performance in both voxel-wise prediction and topological connectivity, which could be very useful in clinical practice.

4 Conclusion

In this paper, we proposed a novel method of automatic vessel segmentation for head and neck CTA images. Specially, our presented work first conduct a classification for semantic prediction using distance weighted BCE loss and dice loss, and a regression for three-axis affinities using constrained malis loss. Second, the final segmentation result can be obtained from the post-processed affinity map and the probability map by intuitive neighbor component connection constraint.

Qualitative and quantitative evaluation results suggest the topological connectivity of the tubular vessel segmentation is largely improved and lead to a superior segmentation than conventional methods such as the well-known nnU-Net. As results, automatic and quantitative analysis of the head and neck vessels is feasible based on the segmentation results. Future works would focus on the segmentation improvement of small vessels as well as those affected by cerebrovascular diseases.

Acknowledgements. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0116400.

References

1. Xu, G., Ma, M., Liu, X., Hankey, G.J.: Is there a stroke belt in china and why? Stroke **44**(7), 1775–1783 (2013)
2. Saxena, A., Ng, E.Y.K., Lim, S.T.: Imaging modalities to diagnose carotid artery stenosis: progress and prospect. Biomed. Eng. Online. **18**(1), 1–23 (2019)
3. Hedblom, A.: Blood vessel segmentation for neck and head computed tomography angiography (2013)
4. Cuisenaire, O., Virmani, S., Olszewski, M.E., Ardon, R.: Fully automated segmentation of carotid and vertebral arteries from contrast-enhanced CTA. In: Medical Imaging 2008: Image Processing, vol. 6914, p. 69143R. International Society for Optics and Photonics (2008)
5. Fan, F., et al.: Rapid vessel segmentation and reconstruction of head and neck angiograms using 3d convolutional neural network. Nat. Commun. **11**(1), 1–12 (2020)
6. Xu, R., Liu, T., Ye, X., Lin, L., Chen, Y.-W.: Boosting connectivity in retinal vessel segmentation via a recursive semantics-guided network. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 786–795. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_76
7. Paetzold, J.C., et al.: clDice-a novel connectivity-preserving loss function for vessel segmentation. In: Medical Imaging Meets NeurIPS 2019 Workshop (2019)
8. Funke, J., et al.: Large scale image segmentation with structured loss based deep learning for connectome reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1669–1680 (2018)
9. Hu, X., Fuxin, L., Samaras, D., Chen, C.: Topology-preserving deep image segmentation. arXiv preprint arXiv:1906.05404 (2019)
10. Qin, Y., et al.: AirwayNet: a voxel-connectivity aware approach for accurate airway segmentation using convolutional neural networks. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 212–220. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_24
11. Kruskal, J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Am. Math. Soc. **7**(1), 48–50 (1956)
12. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
13. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)



Coarse Retinal Lesion Annotations Refinement via Prototypical Learning

Qinji Yu[1], Kang Dang[2], Ziyu Zhou[1], Yongwei Chen[2], and Xiaowei Ding[1,2(✉)]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] VoxelCloud, Inc., Los Angeles, USA
dingxiaowei@sjtu.edu.cn

Abstract. Deep-learning-based approaches for retinal lesion segmentation often require an abundant amount of precise pixel-wise annotated data. However, coarse annotations such as circles or ellipses for outlining the lesion area can be six times more efficient than pixel-level annotation. Therefore, this paper proposes an annotation refinement network to convert a coarse annotation into a pixel-level segmentation mask. Our main novelty is the application of the prototype learning paradigm to enhance the generalization ability across different datasets or types of lesions. We also introduce a prototype weighing module to handle challenging cases where the lesion is overly small. The proposed method was trained on the publicly available IDRiD dataset and then generalized to the public DDR and our real-world private datasets. Experiments show that our approach substantially improved the initial coarse mask and outperformed the non-prototypical baseline by a large margin. Moreover, we demonstrate the usefulness of the prototype weighing module in both cross-dataset and cross-class settings.

Keywords: Prototypical learning · Retina lesion segmentation · Coarse annotation refinement

1 Introduction

Given the growing demand for retinal screening, automatic segmentation for retinal lesions enjoys increasing clinical relevance. By answering the issue of what lesions exist in the image and where they are located, retinal lesion segmentation algorithms assist ophthalmologists in making clinical diagnoses and assessing disease severity [18]. While recent deep-learning approaches have tremendously boosted the retinal lesion segmentation accuracy [4,9,18,19], they often require abundant expert-level-accurate, pixel-wise annotated data, which requires significant time and expense to acquire. Previous studies show that coarse annotations such as circles or ellipses for outlining the lesion area can be six times more efficient than pixel-level annotation [5]. Therefore, it is essential to study novel methodologies tailored for lower-quality coarse annotations.

Q. Yu and K. Dang—Contribute equally to this work.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 239–248, 2022.
https://doi.org/10.1007/978-3-031-21014-3_25

Existing works on exploiting coarse annotations can be categorized into weakly-supervised segmentation [1,10,11,15,16,20,24], and mask refinement [5,21]. Weakly-supervised segmentation methods rely on prior assumptions such as box tightness constraint [16] and image contrast constraint [15] to utilize box-level and image-level coarse annotations. A few high-quality pixel-level retinal lesion datasets such as IDRiD [12] and DDR [8] provide precise lesion boundaries. While successful, weakly-supervised segmentation does not utilize these pre-existing lesion segmentation datasets that provide rich knowledge on lesions' exact appearance and shape. Instead of further developing weakly-supervised segmentation methods, we propose to use such datasets by training an annotation refinement model in a data-driven manner to convert a coarse annotation into a pixel-level segmentation mask. It should be emphasized that our work is significantly different from the weakly-supervised approach, as it is trained in a fully supervised way (instead of weakly-supervised) with coarse annotation and pixel-level ground truth in pairs. Additionally, we note several existing mask refinement methods [5,21] which refine initial coarse masks into more accurate segmentation results; however, they are usually optimized for a particular dataset. In comparison, our method applies the prototype learning paradigm [7,14,17,24] to enhance generalization across different data sets and lesion types. Good generalization is the key to putting the coarse annotation refinement algorithm into practice. For example, we can train the coarse annotation refinement network on a large-scale dataset for once and reuse the trained model on other datasets with less fine-tuning or tweakings.

Particularly, our prototype learning averages the features from the coarse mask region to form a lesion prototype and averages the background features to create a background prototype. A pixel is classified to the lesion class if its corresponding feature vector is more similar to the lesion prototype. Since our prototypical approach generates image-specific prototype to adaptively describe the image itself, it is less sensitive to the intra-class variance and high distribution shifts from different datasets or unseen classes. However, averaging features uniformly may be problematic when the lesion is considerably smaller than the coarse mask, as the resultant lesion prototype becomes dominated by background features. We alleviate this issue by a superpixel-guided prototype weighing module. The module first divides the coarse mask into several superpixels [7] and the prototype for each superpixel is obtained. Each prototype's dis-similarity with the background prototype is then calculated as a weighting factor. The final lesion prototype is the weighted combination of these superpixel-guided sub-prototypes.

Contributions. (1) To the best of our knowledge, our method is the first prototypical approach for the coarse retinal lesion annotation refinement problem. (2) We present a prototype weighing module to solve the problem of the actual lesion being overly small. (3) Experiments demonstrate that the proposed method substantially improved the initial coarse annotation and outperformed non-prototypical mask refinement baselines. It also confirms the superiority of the prototype weighing module in both cross-dataset and cross-class settings.

2 Methods

This section details the proposed coarse annotation refinement method with the overall structure shown in Fig. 1. We assume that there exists a set of image patches and the associated coarse lesion annotations, and our algorithm will convert them into the corresponding high-quality pixel-level annotations.

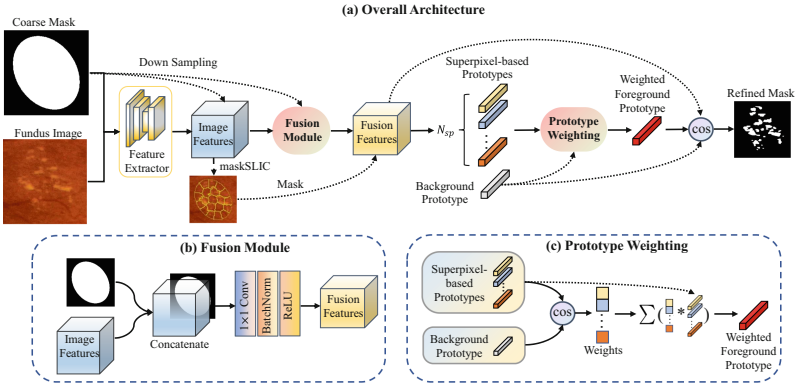


Fig. 1. Framework of our prototype-based coarse annotation refinement network.

2.1 Annotation Refinment via Prototype Learning

Feature Extraction The input to the network is the concatenation of the image patch $I \in \mathbb{R}^{H\times W\times 3}$ and its corresponding coarse lesion annotation $M \in \mathbb{R}^{H\times W\times 1}$. We use a modified U-Net backbone to extract its feature map $F \in \mathbb{R}^{H'\times W'\times C}$. Following [14], we remove the last two upsampling blocks in the U-Net to speed up the calculation. As a result, the resolution of the feature map is 1/4 of the original input. We concatenate the feature map with the down-sampled coarse mask $M' \in \mathbb{R}^{H'\times W'\times 1}$ in the feature channel dimension to further incorporate the coarse annotation prior. To get the final fusion feature map $F' \in \mathbb{R}^{H'\times W'\times C'}$, we adopt a simple 1-layer network with architecture: 1×1 Conv2d+BatchNorm2d+ReLU.

Coarse Prototype Extraction Given the fused feature map, we want to learn representative and well-separated prototype vectors for the lesion region and the background based on the prototypical network. In previous research [14,17,22], the prototypical network condenses the masked object features in an image into a single or few prototypes. A relative simple coarse foreground lesion prototype can be calculated by mask average pooling, as follows:

$$p_{fg} = \frac{\sum_{(x,y)} F'(x,y)\mathbb{1}[M'(x,y)=1]}{\sum_{(x,y)} \mathbb{1}[M'(x,y)=1]}, \tag{1}$$

where (x, y) indexes the spatial locations and $\mathbb{1}(\bullet)$ is an indicator function. In addition, the background prototype is computed by

$$p_{bg} = \frac{\sum_{(x,y)} F'(x,y)\mathbb{1}[M'(x,y) = 0]}{\sum_{(x,y)} \mathbb{1}[M'(x,y) = 0]}, \tag{2}$$

where $p_{fg}, p_{bg} \in \mathbb{R}^{C'}$.

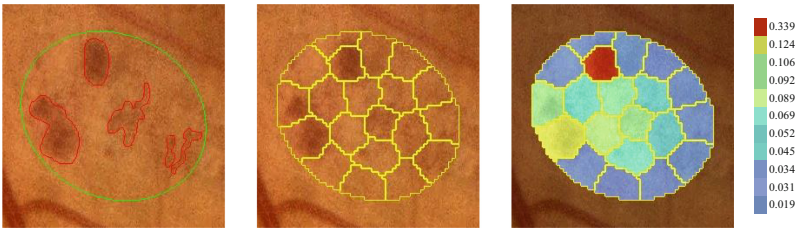
Coarse Annotation Refinement Refinement is done using a non-parametric metric learning method [17]. For each pixel at location (x, y) of the final fusion feature map F', we calculate the distance between its feature vector and the derived prototypes $\mathcal{P} = \{p_{bg}, p_{fg}\}$. Then, we apply the softmax operation over the distances to get the probability map $P_c \in \mathbb{R}^{H' \times W' \times 1}$ and $c \in \{bg, fg\}$. Formally, we have:

$$P_c(x,y) = \frac{\exp(-\alpha \cdot d(F'(x,y), p_c))}{\sum_{p_j \in \mathcal{P}} \exp(-\alpha \cdot d(F'(x,y), p_j))}, \tag{3}$$

where α is the scaling factor fixed at 20.

We train our model end-to-end using the sum of dice loss $\mathcal{L}_{dice}$ and binary cross-entropy loss $\mathcal{L}_{bce}$ between the final probability map P_{fg} and the well-annotated ground truth mask M_{gt}. That is: $\mathcal{L}_{loss} = \mathcal{L}_{dice} + \mathcal{L}_{bce}$.

During testing, for each image patch I and its corresponding coarse lesion annotation M, we obtain a corresponding foreground probability map P_{fg}. When mapping P_{fg} back to the original image space (uncropped full image), some of them will overlap. For each pixel in the overlapping area, we choose the maximum probability of these probability maps as its value. In the end, thresholding is used to get the final refined mask.



(a) Actual lesion region (b) Superpixels generated by maskSLIC (c) Prototype weights heat map

Fig. 2. Illustration of superpixel-guided prototype weighting. Boundaries in red, green, and yellow are actual lesion, coarse lesion, and superpixel regions respectively (Color figure online)

2.2 Superpixel-Guided Prototype Weighting

As shown in Fig. 2 when the actual lesion is relatively small compared to the coarse mask, the coarse foreground prototype defined by Eq. (1) cannot represent the actual lesion features. To reduce the impact of false-positive pixels within the coarse annotation, we divide the initial coarse region into several subregions according to their feature similarity. Concretely, we refer to maskSLIC [6] to aggregate the feature map within the masked region into multiple superpixel clusters. For each superpixel region S_i, we can obtain its corresponding superpixel-based sub-prototype g_i according to Eq. (1). We collect the extracted sub-prototypes and denote them as set $\mathcal{G} = \{g_i\}$, where $i \in 1, 2, \cdots, N_{sp}$ (N_{sp} is the number of superpixels). We compute the cosine distance to measure the similarity between each g_i and p_{bg}:

$$d(g_i, p_{bg}) = 1 - \frac{g_i \cdot p_{bg}}{\|g_i\| \cdot \|p_{bg}\|}. \tag{4}$$

Intuitively, the prototypes dissimilar to the background prototype are more important parts for the final foreground prototype. Therefore, we can derive a weight coefficient for each prototype in set $\mathcal{G}$:

$$w_i = \frac{\exp(\beta \cdot d(g_i, p_{bg}))}{\sum_{g_j \in \mathcal{G}} \exp(\beta \cdot d(g_j, p_{bg}))}, \tag{5}$$

where β is the scaling factor fixed at 10. The final foreground prototype is then given by

$$p_{\text{weighted}} = \sum_{g_i \in \mathcal{G}} w_i \cdot g_i. \tag{6}$$

As shown later, our proposed superpixel weighted prototype p_{weighted} is a more representative foreground prototype that performed better in various experiments.

3 Experiments and Results

3.1 Experimental Setup

Coarse Annotation Generation. There is no public available retinal lesion dataset with paired coarse annotation and pixel-level segmentation mask. To construct such paired dataset, we develop a simple coarse annotation generation method. Firstly, the coarse annotations are simulated from the well-annotated fine masks by applying the following chain of operations: smoothing, dilating, expanding, clustering the connected components using DBSCAN [2] and fitting ellipses to each cluster. Secondly, the fundus image is cropped around each ellipse in the corresponding coarse annotation. Finally, these cropped image patch and coarse annotation pairs are resized to fixed dimensions $H \times W$ for subsequent model training and testing.

Datasets and Evaluation Metrics. We evaluate the proposed methods on publicly IDRiD and DDR datasets, and our real-world private dataset. IDRiD contains 81 fundus images (54 training images, 27 testing images) with pixel-level annotations for hard exudates (EX), hemorrhages (HE), microaneurysms (MA), and soft exudates (SE). Similarly, the testing part of DDR contains 225 fundus images with pixel-level annotations for EX, HE, MA, and SE. Our real-world private testing dataset collects 211 fundus images with pixel-level annotations for drusen (Drus) and pre-retinal hemorrhages (Prh) labeled by two experienced ophthalmologists. To train our annotation refinement network, we collect 32985 training patch pairs (EX:9957, HE:7752, MA:14387, SE:889) from the IDRiD training images using our coarse annotation generation algorithm. We also apply the mask generation algorithm to generate coarse mask for each testing image. To compare different refinement methods, we calculate the Intersection over Union (IoU) between the refined annotation and ground-truth mask.

Baselines. We implement several non-prototypical mask refinement baseline models, taking in an image patch and a coarse mask as the input. In detail, we choose three widely used feature extraction backbones, Res18 [3], HRNet18 [23], and U-Net [13] attached with the coarse mask fusion module to perform feature extraction. The feature extraction process is identical to the one described in Sec.2.1 except for the feature backbone. After that, we attach a 1×1 Conv2d layer as a binary classifier to obtain a refined segmentation score map.

Implementation Details. For the prototypical methods, we set the superpixel number $N_{sp} = 20$. All training patch pairs are resized to 256×256 and augmented by RandomShiftScaleRotate, RandomBlur, and RandomBrightnessContrast. All models are implemented by PyTorch and trained from scratch using Adam optimizer with a batch size of 64 for 120 epochs. The initial learning rate is 10^{-4} and reduces according to ReduceLROnPlateau strategy.

3.2 Results

Same-Dataset Experiments. We train the proposed coarse annotation refinement network using all four lesion types on IDRiD and evaluate the performance on the IDRiD testing set. As shown in Table 1, "Initial Coarse" denotes the IoU of actual lesion region versus coarse annotation region. Our prototypical method improves the initial coarse mask considerably. It also consistently obtains better refinement performance than the non-prototypical baselines on all four lesion classes in terms of IoU score, with average IoU score improving by more than 5.2%. This experiment demonstrates our advantages when training and testing images are from the same dataset.

Cross-Dataset and Cross-Class Experiments. We directly evaluate the performance on the DDR testing set and our real-world private dataset using

Table 1. The image-level average IoU (%) and its standard deviations (%) of on IDRiD."w/ superpixel" means with superpixel-guided prototype weighing.

Methods	MA	SE	EX	HE	Average
Initial Coarse	9.6 (2.6)	49.3 (10.4)	15.0 (5.2)	33.2 (8.6)	26.8 (6.7)
Res18 [3]	73.9 (7.3)	68.4 (14.2)	54.1 (9.4)	62.6 (10.9)	64.7 (10.5)
HRNet18 [23]	79.1 (7.5)	78.1 (5.3)	56.8 (9.1)	64.8 (11.8)	69.7 (8.4)
U-Net [13]	77.6 (6.8)	75.6 (12.6)	58.9 (8.9)	67.9 (11.4)	70.0 (9.9)
Our methods	**84.2** (6.2)	**80.7** (9.2)	65.3 (8.1)	69.9 (13.9)	75.0 (9.4)
w/ superpixel	84.1 (6.2)	79.6 (9.9)	**65.9** (8.7)	**71.1** (11.2)	**75.2** (9.0)

models trained on the IDRiD dataset without further fine-tuning. As shown in Table 2, our method exceeds the U-Net baseline by 4.3% on both DDR and private datasets. For DDR, our superpixel weighted prototype performs better for all lesion types compared to the non-weighted prototype. Similarly, the weighted prototype is notably better than the non-weighted one on the private dataset, especially for the class Prh (56.3% → 62.6%). Overall, we see a general trend that our model can generalize well to new datasets or unseen classes.

Table 2. The image-level average IoU (%) and its standard deviations (%) on DDR and our real-world private dataset having 2 unseen classes.

			DDR				Private	
Methods	MA	SE	EX	HE	Average	Drus	Prh	Average
Initial Coarse	6.9 (3.4)	32.3 (10.8)	14.8 (8.8)	23.1 (11.4)	19.3 (8.6)	33.6 (19.1)	49.7 (13.0)	41.6 (16.0)
Res18 [3]	56.3 (12.4)	67.6 (13.6)	52.3 (13.2)	54.4 (13.3)	57.6 (13.1)	43.9 (16.6)	52.9 (15.2)	48.4 (15.9)
HRNet18 [23]	65.1 (13.5)	68.2 (12.6)	54.2 (13.9)	58.5 (12.9)	61.5 (13.3)	44.2 (21.1)	57.5 (16.4)	50.9 (18.8)
U-Net [13]	60.9 (13.2)	70.2 (15.0)	55.4 (12.8)	59.8 (12.7)	61.6 (13.4)	45.2 (20.9)	57.9 (16.4)	51.6 (18.6)
Our methods	68.8 (12.0)	71.2 (16.9)	58.4 (12.7)	61.3 (15.7)	64.9 (14.3)	47.9 (23.4)	56.3 (20.7)	52.1 (22.1)
w/ superpixel	**69.8** (11.9)	**72.0** (17.5)	**58.9** (12.4)	**62.9** (14.5)	**65.9** (14.1)	**49.1** (23.0)	**62.6** (15.3)	**55.9** (19.1)

Coarse Mask Reduction Factors. Since ophthalmologists tend to draw a single rough ellipse to cover several unconnected lesion regions, we simulate the process by setting different reduction factors to the DBSCAN clustering algorithm. Actually, the number of the generated ellipses is the number of connected lesion regions divided by the reduction factor. In other words, with a higher reduction factor, the generated coarse mask will be more coarse. As shown in Fig. 3, although the refinement performance of all methods degrades as the reduction factor ranges from 1.0 to 2.0, our prototypical method has less degradation compared to the U-Net baseline. It implies our method is more robust against coarser annotations.

Visual Results. Figure 4 presents some visualization of refinement results. Despite the vast variation in lesion scales, colors, and low contrast to surrounding regions, the first three rows show our proposed superpixel weighted prototype approach generates the most accurate lesion boundary. The last row shows a failure case when the coarse mask contains two distinct lesion classes, EX and Drus, at the same time.

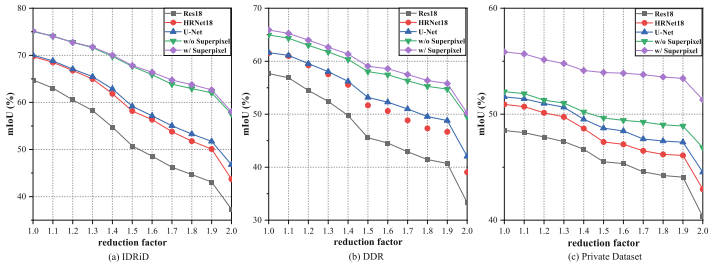


Fig. 3. Refinement performance under different reduction factors.

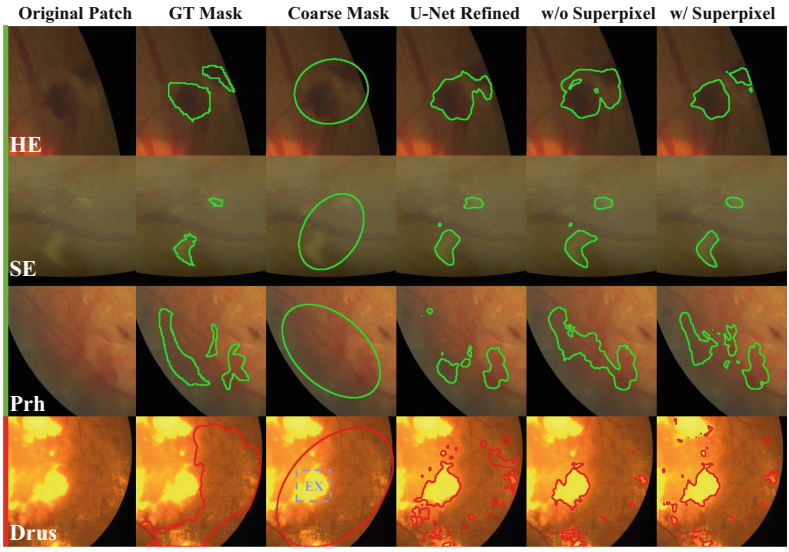


Fig. 4. Visualization of refinement results on four categories of lesions.

4 Conclusions

This paper proposes a novel prototype-based network to convert a coarse annotation into a pixel-level segmentation mask. The proposed network first extracts the lesion and background prototypes and labels the image pixel as the lesion

class if its feature is more similar to the lesion prototype. A superpixel-guided prototype weighing module is then proposed to tackle the issue of the actual lesion being overly small compared to the coarse mask. On the IDRiD dataset, our model outperformed non-prototypical baselines by a large margin. Extensive experiments on DDR and our real-world private dataset also demonstrate the proposed model enjoys better generalizability to new datasets and some unseen lesion classes.

References

1. Chu, T., Li, X., Vo, H.V., Summers, R.M., Sizikova, E.: Improving weakly supervised lesion segmentation using multi-task learning. In: Medical Imaging with Deep Learning, pp. 60–73. PMLR (2021)
2. Ester, M., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol. 96, pp. 226–231 (1996)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
4. Huang, S., Li, J., Xiao, Y., Shen, N., Xu, T.: RTNet: relation transformer network for diabetic retinopathy multi-lesion segmentation. IEEE Trans. Med. Imaging. **41**, 1596–1607 (2022)
5. Huang, Y., et al.: Automated hemorrhage detection from coarsely annotated fundus images in diabetic retinopathy. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1369–1372. IEEE (2020)
6. Irving, B.: maskslic: regional superpixel generation with application to local pathology characterisation in medical images. arXiv preprint arXiv:1606.09518 (2016)
7. Li, G., Jampani, V., Sevilla-Lara, L., Sun, D., Kim, J., Kim, J.: Adaptive prototype learning and allocation for few-shot segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8334–8343 (2021)
8. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Inf. Sci. **501**, 511–522 (2019)
9. Liu, Q., Liu, H., Liang, Y.: M2MRF: Many-to-many reassembly of features for tiny lesion segmentation in fundus images. arXiv preprint arXiv:2111.00193 (2021)
10. Liu, X., et al.: Weakly supervised segmentation of covid19 infection with scribble annotation on CT images. Pattern Recogn. **122**, 108341 (2022)
11. Playout, C., Duval, R., Cheriet, F.: A novel weakly supervised multitask architecture for retinal lesions segmentation on fundus images. IEEE Trans. Med. Imaging **38**(10), 2434–2444 (2019)
12. Porwal, P., et al.: IDRiD: diabetic retinopathy-segmentation and grading challenge. Med. Image Anal. **59**, 101561 (2020)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Tang, H., Liu, X., Sun, S., Yan, X., Xie, X.: Recurrent mask refinement for few-shot medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3918–3928 (2021)

15. Tang, Y., et al.: Weakly-supervised universal lesion segmentation with regional level set loss. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 515–525. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_48
16. Wang, J., Xia, B.: Bounding box tightness prior for weakly supervised image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 526–536. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_49
17. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: PANET: few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9197–9206 (2019)
18. Wei, Q., et al.: Learn to segment retinal lesions and beyond. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 7403–7410. IEEE (2021)
19. Yan, Z., Han, X., Wang, C., Qiu, Y., Xiong, Z., Cui, S.: Learning mutually local-global u-nets for high-resolution retinal lesion segmentation in fundus images. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 597–600. IEEE (2019)
20. Yang, L., et al.: BoxNet: deep learning based biomedical image segmentation using boxes only annotation. arXiv preprint arXiv:1806.00593 (2018)
21. Yang, Y., Wang, Z., Liu, J., Cheng, K.T., Yang, X.: Label refinement with an iterative generative adversarial network for boosting retinal vessel segmentation. arXiv preprint arXiv:1912.02589 (2019)
22. Yu, Q., Dang, K., Tajbakhsh, N., Terzopoulos, D., Ding, X.: A location-sensitive local prototype network for few-shot medical image segmentation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 262–266. IEEE (2021)
23. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation (2020)
24. Zhang, G., et al.: RefineMask: towards high-quality instance segmentation with fine-grained features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6861–6869 (2021)



Nuclear Segmentation and Classification: On Color and Compression Generalization

Quoc Dang Vu, Robert Jewsbury(✉), Simon Graham, Mostafa Jahanifar, Shan E. Ahmed Raza, Fayyaz Minhas, Abhir Bhalerao, and Nasir Rajpoot(✉)

Tissue Image Analytics Centre, University of Warwick, Coventry, UK
{quoc-dang.vu,rob.jewsbury,n.m.rajpoot}@warwick.ac.uk

Abstract. Since the introduction of digital and computational pathology as a field, one of the major problems in the clinical application of algorithms has been the struggle to generalize well to examples outside the distribution of the training data. Existing work to address this in both pathology and natural images has focused almost exclusively on classification tasks. We explore and evaluate the robustness of the 7 best performing nuclear segmentation and classification models from the largest computational pathology challenge for this problem to date, the CoNIC challenge. We demonstrate that existing state-of-the-art (SoTA) models are robust towards compression artifacts but suffer substantial performance reduction when subjected to shifts in the color domain. We find that using stain normalization to address the domain shift problem can be detrimental to the model performance. On the other hand, neural style transfer is more consistent in improving test performance when presented with large color variations in the wild.

Keywords: Robustness · Nuclei segmentation and classification

1 Introduction

The spatial arrangement and morphology of nuclei are important signatures for identifying disease [5,20]. In cancer, bio-markers such as tumor-infiltrating-lymphocytes [19] or Programmed death-ligand 1 (PD-L1) combined positive score for response to immunotherapy [6] have been found to be highly correlated with patient survival. In spite of their supposed effectiveness, their adoption and use in clinical settings remain limited due to the high degree of inter and intra observer variation [27] of the derived scores.

To overcome this limitation, many machine learning challenges [13,24] have been proposed in digital pathology to further the innovation of automating the identification and classification of nuclei instances in Haematoxylin & Eosin (HE) stained samples. The most recent and perhaps the largest challenge of its kind to date, with 704 participants in 96 teams, the Colon Nuclei Identification and Counting Challenge 2022 (CoNIC) [10], was able to identify several powerful deep-learning-based methods for these tasks [1–3,14,17,26,29]. These methods

Q. D. Vu and R. Jewsbury—Joint First Authors Contributed Equally.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 249–258, 2022.
https://doi.org/10.1007/978-3-031-21014-3_26

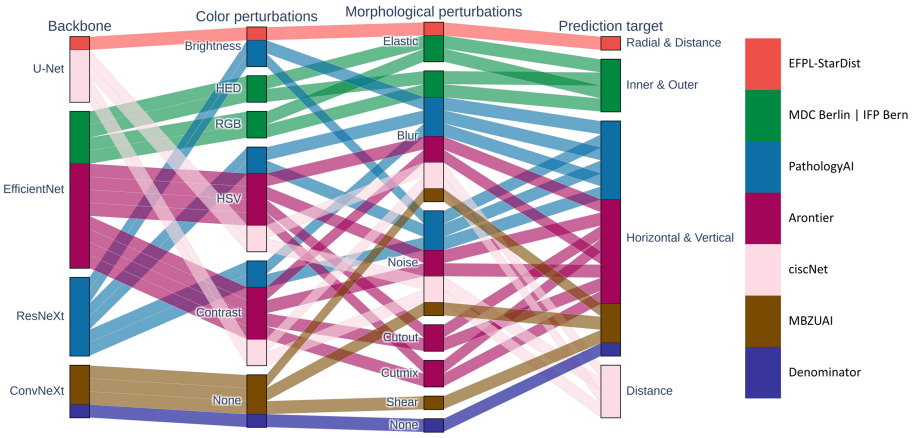


Fig. 1. Summary of assessed segmentation methods from the CoNIC challenge. Significant aspects including the backbone network family, training augmentations related to color and compression (morphology) and prediction target are shown.

achieved competitive performance even on an unseen, large scale cohort. While the results are promising, it is as yet unclear whether these methods are sufficient to achieve generalization across millions of nuclei in the wild for subsequent downstream analyses.

This question about generalization in the wild is somewhat synonymous with investigating the robustness of models with respect to the shift of input distribution (or domain shift). So far in digital pathology, such investigations have been limited mostly to patch-level classification tasks [21,28]. Concurrent to these developments, in natural images, attempts to see how deep neural networks (DNNs) fare against various input alterations have uncovered several important insights. To name a few, existing DNNs have a shape and texture bias [8], this limits the model to reason further about the geometries of their learning targets. Interestingly, this limitation relates to the commonly chosen 3×3 kernel [4].

With this perspective, it is important to gain better insights about the existing state-of-the art (SoTA) *segmentation-based* methods for identification and classification of nuclear instances. Specifically, in this paper, we investigate how the performance of the top performing approaches in the CoNIC challenge are affected by perturbations to the input. We focus on two forms of domain shift: artifacts introduced by compressing whole slide images and alteration of image colors. We empirically show that:

- Upon the finalization of training, each approach has a virtual center of color distribution that is different from that of the training set;
- Common augmentations employed during training may not improve robustness against color and morphological perturbations as practitioners might expect;
- Stain normalization can be more detrimental to performance than beneficial and needs to be employed very carefully.

2 Methodology

2.1 Robustness Investigations

As noted by [11], there is no widely accepted definition for robustness. In this paper, we consider 2 aspects which determine how a sample domain becomes *shifted* compared to that of another sample and investigate how the performance of segmentation-based methods is affected in such cases: a) *Compression related artifacts* and b) *Color variation* between tissue, organs, diseases, scanning devices and staining protocols.

2.2 Segmentation-based Methods

We explored the top 7 performing models in the CoNIC challenge that use a segmentation-based approach: Denominator [14], MBZUAI [2], ciscNet [3], Arontier [1], PathologyAI [29], MDC Bern - IFP Berlin [17] and EFPL - Stardist [26]. We summarize various aspects of these approaches that can influence how the models fare against the alterations under investigation in Fig. 1. All methods follow the encoder-decoder architecture paradigm and only EFPL - Stardist model was trained from scratch.

2.3 Evaluation Metrics

The CoNIC challenge utilized a variant of panoptic quality (PQ) [12] for multi-class problems called mPQ^+ to measure the performance of segmentation and classification of nuclei instances. At $\text{IoU}(x_t^a, y_t^a) > a$, the PQ_t^a for each type of nucleus t is defined as

$$\mathcal{PQ}_t^a = \underbrace{\frac{|TP_t^a|}{|TP_t^a| + \frac{1}{2}|FP_t^a| + \frac{1}{2}|FN_t^a|}}_{\text{Detection Quality(DQ)}} \times \underbrace{\frac{\sum_{(x_t^a, y_t^a) \in TP} IoU(x_t^a, y_t^a)}{|TP_t^a|}}_{\text{Segmentation Quality(SQ)}} \tag{1}$$

Here, x_t denotes a ground truth GT (GT) instance, y_t denotes a predicted instance, and IoU denotes intersection over union. A unique pairing between a GT and predicted instance is derived when setting the threshold $a \geq 0.5$ for $\text{IoU}(x^a, y^a)$, or by using Hungarian matching for $a < 0.5$. This matching splits all available instances of type t within *an image* into matched pairs (TP), unmatched GT instances (FN) and unmatched predicted instances (FP). In medical images, IoU is a harsh criterion for problems that contain lots of small objects such as nuclei [15]. However, for downstream analysis, often all predicted nuclei are utilized rather than just a subset being above a certain size. We extend mPQ^+ by taking its average across a range $D = \{a \in \mathbb{R}; 0.0 \leq a \leq 0.5\}$. This is synonymous to calculating the area under the curve (AUC), thus we define

$$m\mathcal{PQ}^+ AUC = \int_0^{0.5} \frac{1}{T} \sum_t^{T=6} \mathcal{PQ}^+{}_t^a \, \mathrm{d}a \tag{2}$$

To reduce the computation cost, we sampled α with a step of 0.05 and used the trapezoidal rule to obtain the final results.

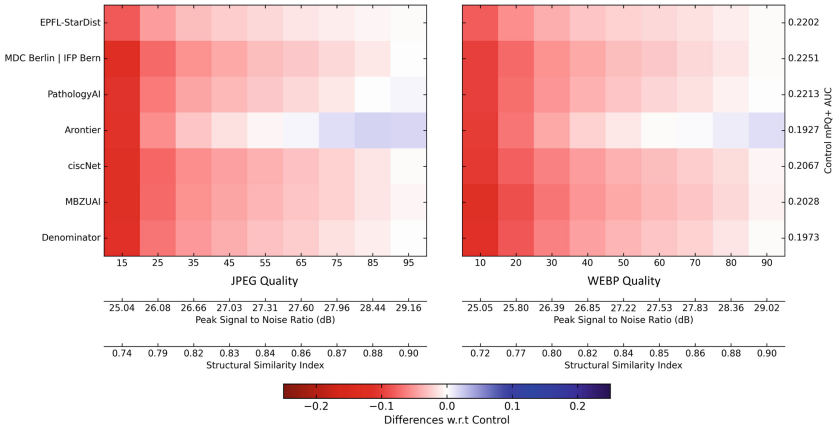


Fig. 2. Changes in the $m\mathcal{PQ}^{+}AUC$ of SoTA segmentation-based methods when varying the *quality* parameter in JPEG or WEBP compression.

3 Experimental Results

3.1 Dataset and Comparison Settings

Training Set: Participants in the CoNIC challenge utilized the development set of the Lizard dataset [9] for training, validating and selecting their models. In total, there are 431,913 unique nuclei belonging to 6 nuclear categories originating from 5 centers.

Testing Set: The challenge test set includes images taken from 12 different centers, where 11 of these are completely unseen in the development set. Data from these 11 centers acted as the external test set in the original Lizard publication [9]. The remaining center comprises of additional data extracted from a center that was already considered during training, but from an independent cohort. In total, there exists 103,150 unique nuclei belonging to 6 nuclear categories.

Comparison Settings: Images from both the Lizard dataset and the testing set were prepared into sets of 256×256 image patches [10]. To facilitate the discovery for new insights, the selected teams from CoNIC provided a single model that was trained, validated and selected based on a fixed split of the Lizard dataset, provided by the organizers (80/20 for training/validation). Analyses in this paper are based on these models rather than the original submissions to the challenge.

Original, Control and Assessed: We denote **Original** as the un-altered version of the test set. The whole slide images from which these patches originated could be at different resolutions and compressed by different methods, we aimed to align their information as much as possible before further compressing the data. When shifting the color domain of the test set, we found that neural style transfer [7] (neural stylization) preserves the local characteristics of the images the best at high resolution. We first created a **Control** version of the test set

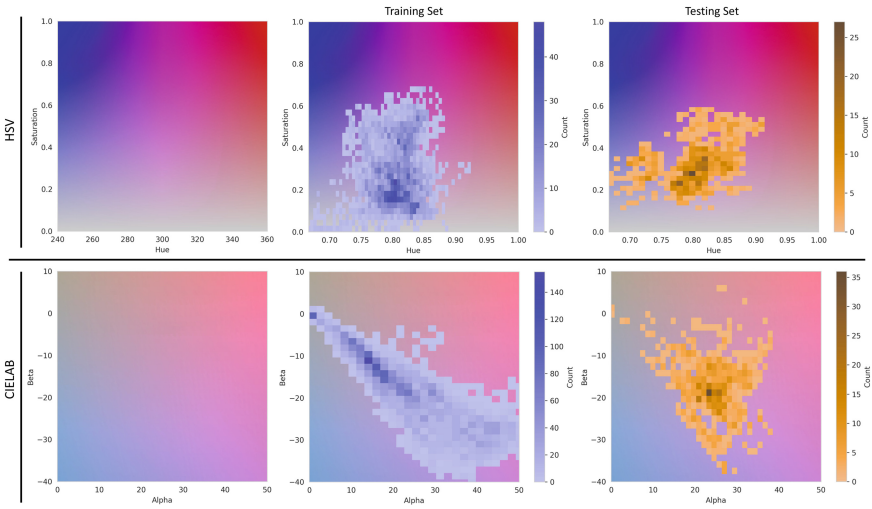


Fig. 3. Color distribution of image patches within the CoNIC training and testing set using kernel density estimation in HSV and CIELAB color space.

by up-scaling the **Original** version to super-resolution via ESRGAN [25] before resizing it back to original resolution. By perturbing this process at various stages, we obtained samples for assessing compression (compressing on **Control**) and shifting color domains (at super-resolution). We denote these versions of the test set as the **Assessed**.

3.2 Compression Perturbation

We evaluated two major lossy compression methods: JPEG and WEBP. We show how the performance of the selected methods varies across a range of compression qualities in Fig. 2. Intuitively, a majority of the methods experience a decrease in performance the more the images are compressed. Interestingly, the Arontier team's model had noticeable performance gain at some compression levels compared to the **Control** results. From Fig. 1, we hypothesise this could be due to the cutout [22] and cutmix [18] augmentations they employed during training as they would introduce compression-like artifacts. None of the other teams under evaluation employed these augmentation techniques.

3.3 Color Perturbation

As each dataset occupies a region within a color space, visualizing how the model performances vary in such a color space is synonymous with approximating the theoretical robustness range of such a model (obtained from a specific training set) with respect to possible samples in the wild. Thus, by identifying how these changes relate to the color distribution of the training and testing set, may provide insights on which aspect of these methods are important for generalization.

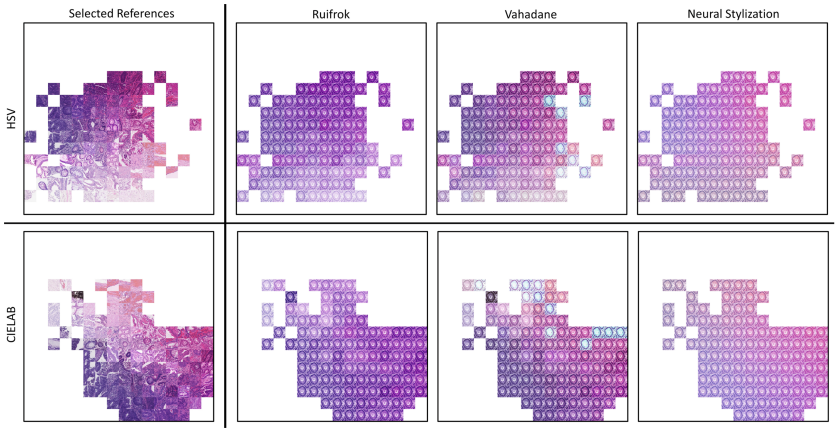


Fig. 4. Selected reference images within the training set and a test sample that has it color pulled toward the reference at the same location in HSV and CIELAB color space. The color ranges in both color space are the same as in Fig. 3.

To ascertain that any such observations are not restricted to a specific color space, we repeat our experiments in both the HSV and CIELAB color spaces.

Color Domains. We represent the color of each image patch by the mean of its pixel values within that color space. With these values, we establish how the entire training and test set is distributed in terms of Hue (H) and Saturation (S) for HSV or Alpha (A) and Beta (B) for CIELAB in Fig. 3 via kernel density estimation. Here, we observe that the center of the test set in CIELAB is noticeably different from that of the training set while they heavily overlap in HSV.

Color Sampling and Artificial Domain Shifting: Having defined the color domains to be investigated qualitatively in Fig. 3, further investigations required us shifting the test set toward a color value in such domain while preserving the image compositions. We achieved this by using stain-normalization or by neural stylization. Both techniques however, require a choice of reference image which represents the distribution of the domain we want to shift our target image towards.

To obtain them, we started by simplifying each color space. We restricted the value ranges of the axis within each color space such that they encapsulate the training and testing set distributions. We achieved this by setting $H \in [240, 360]$, $S \in [0.0, 1.0]$ or $A \in [0, 50]$, $B \in [-40, 10]$ based on Fig. 3. Afterward, along each axis and within these value ranges, we quantized them into 16 steps, thus creating a total of 256 sampling points (16×16) in HSV or CIELAB respectively. In other words, we obtained 256 samples within the training set which were closest to these 256 colors in each color space. Specifically, by setting the $Value$ (V) and $Luminance$ (L) to the mean value of the entire training set, we correspondingly obtained unique images within the training set that were the closest to each

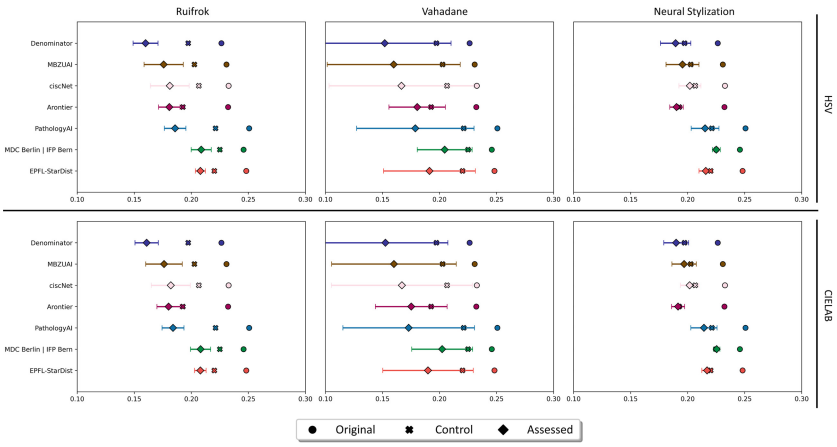


Fig. 5. Changes in the $m\mathcal{PQ}^{+}AUC$ when we artificially shifted the color of the test set using various methods. The box plots are the means and standard deviations.

sampling point. Although there were 256 sampling points, the nearer towards the edge of the training distribution the sampling points became the more likely they shared the same selected images. In such cases, we only retained the ones that have smallest distance within the color space, thus obtaining 101 references for HSV and 108 for CIELAB.

Finally, because different color alteration methods behave differently, in order to ensure the validity of the observed phenomena, we also assessed the results obtained by using three different methods: Ruifrok [16] and Vahadane [23] stain-normalization and neural style transfer [7]. We provide the sampled references and a standard testing sample with their altered color in Fig. 4. Notice that each image in the figure represents a quantized color range as describe above. Additionally, for some reference images, Vahadane did not produce physically realistic samples (the teal colored images). For such cases, we excluded these from our subsequent quantitative analyses detailed in Fig. 5.

Reflected Performances: For each color alteration method, we compute the average and standard deviation of $mPQ^{+}AUC$ across all versions of the testing set in each color space and plot them in Fig. 5. We observed that **Original** results are higher than that of the **Control** results. While all methods suffered to varying degrees from the color shift, MDC Bern | IFP Bern and Arontier were affected the least. Across the two color spaces and three color shift techniques, they had the overall smallest standard deviation and reduction in performance. Interestingly, we found stain-normalization to be detrimental to the model performances in general. However, neural stylization was found to be the most stable (having the smallest deviation).

We further show how these performance changes relate to the choice of reference images for the MDC Bern | IFP Bern, Arontier and Pathology AI teams in Fig. 6. In line with our intuition, moving the color of the testing set toward

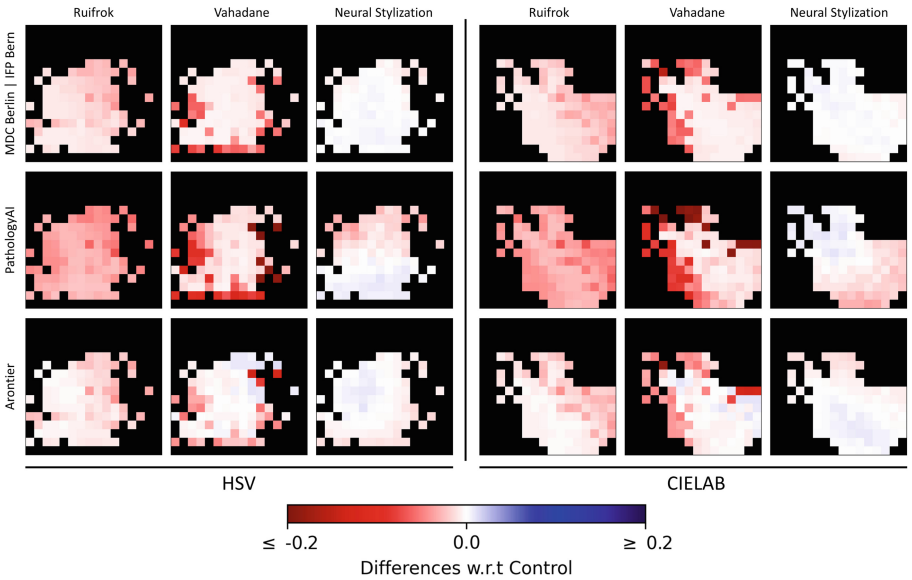


Fig. 6. Difference in the $m\mathcal{PQ}^{+}AUC$ between the **Control** and the test set where its color was shifted w.r.t each sampled reference in Fig. 4.

the center of the training distribution can potentially improve performance, as reflected by PathologyAI. The reference images that demonstrated the smallest reduction in performance (in the case of stain-normalization) or improvement (with neural style-transfer) are found in the center of the training distribution. However, intriguingly, Arontier gained performance when the testing set moved toward areas that are not heavily populated within the training set. Meanwhile, MDC Bern | IFP Bern performed well across the range of sampled color references, with the exception of those being near the training distribution edges. We surmised this is the reason for its large deviation as seen in Fig. 5. Finally, in the context of this task, Vahadane normalization in particular is shown to be highly dependent on the choice of reference image i.e. being at the center of the training distribution, regardless of the models.

4 Conclusions

We empirically show that SoTA segmentation-based methods are quite robust against compression artifacts. If lossy compression is required, default parameters for JPEG (quality 75) and WEBP (quality 80) are sufficient. However, these models are highly volatile with respect to color variations. We find evidence that stain normalization improves model generalization. Specifically, regardless of the choice of method and color space, we show that stain normalization is more likely to degrade performance than improve it. If altering the color (or stain normalization) of the dataset is required, we suggest using neural style transfer [7] as

we found it to provide the most consistent gain in performance. If using stain-normalization, reference images should be from the central area of the training distribution as we found this avoids the outlier regions which resulted in substantial performance loss.

Finally, this work demonstrates that the generalization problem is a concern inspite of the improved performance of computational pathology models in nuclear classification and segmentation tasks achieved in recent years. Our findings highlight the need for further investigations into the mechanism behind these underlying problems.

References

1. Ahn, H., Hong, Y.: Class-controlled copy-paste based cell segmentation for conic challenge. bioRxiv (2022)
2. Azzuni, H., Ridzuan, M., Xu, M., Yaqub, M.: Color space-based hover-net for nuclei instance segmentation and classification. arXiv preprint arXiv:2203.01940 (2022)
3. Böhland, M., et al.: CISCNet-a single-branch cell instance segmentation and classification network. arXiv preprint arXiv:2202.13960 (2022)
4. Ding, X., Zhang, X., Zhou, Y., Han, J., Ding, G., Sun, J.: Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs. arXiv preprint arXiv:2203.06717 (2022)
5. Elston, C., Ellis, I.: Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. **19**(5), 403–410 (1991). https://doi.org/10.1111/j.1365-2559.1991.tb00229.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2559.1991.tb00229.x
6. Fusi, A., et al.: PD-L1 expression as a potential predictive biomarker. Lancet Oncol. **16**(13), 1285–7 (2015)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. CoRR abs/1508.06576 (2015). arxiv.org:1508.06576
8. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. CoRR abs/1811.12231 (2018). arxiv.org:1811.12231
9. Graham, S., et al.: Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification. arXiv preprint arXiv:2108.11195 (2021)
10. Graham, S., et al.: Conic: Colon nuclei identification and counting challenge 2022. CoRR abs/2111.14485 (2021). arxiv.org:abs/2111.14485
11. Hendrycks, D., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. CoRR abs/2006.16241 (2020). arxiv.org:abs/2006.16241
12. Kirillov, A., He, K., Girshick, R.B., Rother, C., Dollár, P.: Panoptic segmentation. CoRR abs/1801.00868 (2018). arxiv.org:abs/1801.00868
13. Kumar, N., et al.: A multi-organ nucleus segmentation challenge. IEEE Trans. Med. Imaging **39**(5), 1380–1391 (2019)
14. Li, J., Wang, C., Huang, B., Zhou, Z.: Convnext-backbone HoverNet for nuclei segmentation and classification. arXiv preprint arXiv:2202.13560 (2022)
15. Maier-Hein, et al.: Metrics reloaded: Pitfalls and recommendations for image analysis validation (2022). arxiv.org:abs/2206.01653

16. Ruifrok, A.C., Johnston, D.A., et al.: Quantification of histochemical staining by color deconvolution. Anal. Quant. Cytol. Histol. **23**(4), 291–299 (2001)
17. Rumberger, J.L., Baumann, E., Hirsch, P., Kainmueller, D.: Panoptic segmentation with highly imbalanced semantic labels. arXiv preprint arXiv:2203.11692 (2022)
18. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)
19. Shield, K.D., et al.: The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012. CA: Cancer J. Clin. **67**(1), 51–64 (2017)
20. Solis, L.M., et al.: Histologic patterns and molecular characteristics of lung adenocarcinoma associated with clinical outcome. Cancer **118**(11), 2889–2899 (2012)
21. Stacke, K., Eilertsen, G., Unger, J., Lundström, C.: Measuring domain shift for deep learning in histopathology. IEEE J. Biomed. Health Inform. **25**(2), 325–336 (2020)
22. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
23. Vahadane, A., et al.: Structure-preserving color normalization and sparse stain separation for histological images. IEEE Trans. Med. Imaging **35**(8), 1962–1971 (2016)
24. Verma, R., et al.: MoNuSAC 2020: a multi-organ nuclei segmentation and classification challenge. IEEE Trans. Med. Imaging **40**(12), 3413–3423 (2021). https://doi.org/10.1109/TMI.2021.3085712
25. Wang, X., et al.: ESRGAN: enhanced super-resolution generative adversarial networks. CoRR abs/1809.00219 (2018). arxiv.org:abs/1809.00219
26. Weigert, M., Schmidt, U.: Nuclei segmentation and classification in histopathology images with stardist for the conic challenge 2022. arXiv preprint arXiv:2203.02284 (2022)
27. Yamaguchi, T., et al.: Inter-observer agreement among pathologists in grading the pathological response to neoadjuvant chemotherapy in breast cancer. Breast Cancer **25**(1), 118–125 (2018)
28. Yamashita, R., Long, J., Banda, S., Shen, J., Rubin, D.L.: Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. IEEE Trans. Med. Imaging **40**(12), 3945–3954 (2021)
29. Zhang, W.: Conic solution. arXiv preprint arXiv:2203.03415 (2022)



Understanding Clinical Progression of Late-Life Depression to Alzheimer's Disease Over 5 Years with Structural MRI

Lintao Zhang[1,2], Minhui Yu[2], Lihong Wang[3], David C. Steffens[3], Rong Wu[4], Guy G. Potter[5(✉)], and Mingxia Liu[2(✉)]

[1] School of Information Science and Engineering, Linyi University, Shandong, China
[2] Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill 27599, USA
mxliu@med.unc.edu
[3] Department of Psychiatry, University of Connecticut School of Medicine, University of Connecticut, Farmington, CT, USA
[4] Connecticut Convergence Institute for Translation in Regenerative Engineering, University of Connecticut Health, Farmington, USA
[5] Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, NC, USA
guy.potter@duke.edu

Abstract. Previous studies have shown that late-life depression (LLD) may be a precursor of neurodegenerative diseases and may increase the risk of dementia. At present, the pathological relationship between LLD and dementia, in particularly Alzheimer's disease (AD) is unclear. Structural MRI (sMRI) can provide objective biomarkers for the computer-aided diagnosis of LLD and AD, providing a promising solution to understand the clinical progression of brain disorders. But few studies have focused on sMRI-based predictive analysis of clinical progression from LLD to AD. In this paper, we develop a deep learning method to predict the clinical progression of LLD to AD up to 5 years after baseline time using T1-weighted structural MRIs. We also analyze several important factors that limit the diagnostic performance of learning-based methods, including data imbalance, small-sample-size, and multi-site data heterogeneity, by leveraging a relatively large-scale database to aid model training. Experimental results on 308 subjects with sMRIs acquired from 2 imaging sites and the publicly available ADNI database demonstrate the potential of deep learning in predicting the clinical progression of LLD to AD. To the best of our knowledge, this is among the first attempts to explore the complex pathophysiological relationship between LLD and AD based on structural MRI using a deep learning method.

Keywords: Late-life depression · Cognitive impairment · Alzheimer's disease · Structural MRI

1 Introduction

Depression, cognitive impairment (CI) and dementia in late life are very common disorders worldwide that affect people's late life. With the expansion of

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 259–268, 2022.
https://doi.org/10.1007/978-3-031-21014-3_27

social, medical and economic burden, the number of patients suffering from these diseases is increasing exponentially. Studies have shown that there is a close relationship between depression, cognitive impairment in late life and progressive dementia, especially Alzheimer's disease (AD) [1–8]. Even though the pathologic progression mechanisms of late-life depression (LLD) and AD remain unclear, these two entities are now thought to be related, *i.e.*, LLD may be a prodromal symptom of neurodegenerative diseases [1]. Currently, there is a lack of objective biomarkers to help us understand the clinical progression from LLD to AD, thus greatly affecting the research results when trying to further explore the relationship between them [9].

At present, the diagnosis of depression is mainly based on clinical interview and standardized assessments (rather than using clinically recognized biomarkers), which may be subjective and biased. It is therefore relevant to examine imaging biomarkers for the progression prediction of LLD to AD, so as to promote the early intervention and effective treatment of the related diseases. Neuroimaging provides a promising solution for understanding the pathophysiological relationship between LLD and AD, and will support biomarker-driven diagnosis. In particular, structural magnetic resonance imaging (sMRI) provides a non-invasive solution to objectively quantify pathological changes in the brain that lead to significant mental illness. Increasing evidence in the literature has shown that changes in white matter and gray matter are directly related to depressive symptoms [2,10,11]. A previous study has shown that LLD patients developing CI and dementia can be discriminated from LLD patients remaining cognitively stable with good accuracy based on baseline structural MRI [3]. In [12], deep learning model (*i.e.*, VGG16) and an SVM classifier are used for feature extraction and classification, respectively, to automatically detect major depressive disorder (MDD) based on sMRI scans. Joko *et al.* [5] suggest that the method of comparing hippocampal atrophy may be useful in distinguishing AD, CI, MDD, and normal controls. However, these machine learning algorithms generally rely on hand-crafted MRI features for automated classification and prediction of AD, CI, and MDD. A recent study [13] employs a deep learning model to extract sMRI features for depression symptom factors regression, demonstrating the potential of deep learning in finding objective MRI biomarkers for depression analysis. Unfortunately, few studies have focused on exploring the clinical progression from LLD to AD using structural MRI data.

In this paper, we developed a deep learning method to predict individuals who progressed to AD within 5 years using baseline T1-weighted MRI. We also analyzed several important factors that might limit the prediction performance of our method, by leveraging a relatively large-scale database to aid model training. We conducted imaging data analyses on 308 subjects with sMRIs acquired from 2 imaging sites and 360 subjects from the publicly available ADNI database [14], with results demonstrating the potential of deep learning in predicting the clinical progression of LLD to AD. To our knowledge, this is among the first attempts to explore the complex pathophysiological relationship between LLD and AD based on sMRI using deep learning methods.

2 Materials and Proposed Method

2.1 Materials and Image Pre-processing

This study was performed in line with the principles of the Declaration of Helsinki and approved by the Duke University Institutional Review Board. All participants were enrolled in the Neurocognitive Outcomes of Depression in the Elderly (NCODE) study [15] and the Neurobiology of Late-life Depression (NBOLD) study [16]. The T1-weighted MRI scans in NCODE were acquired using a 3D axial TURBOFLASH sequence with TR/TE=22/7 ms, flip angle = 25°, 100 Hz/pixel bandwidth, a 256×256 matrix, a 256 mm diameter field-of-view, 160 slices with a 1 mm slice thickness and Nex=1 (no signal averaging), yielding an image with 1 mm cubic voxels. All sMRI scans in NBOLD were acquired using a Skyra 3T scanner (Siemens, Erlangen, Germany) with 32 surface coils located at Olin Neuropsychiatric Research Center (ONRC). Five high resolution axial T1-weighted magnetization-prepared rapid gradient-echo (MPRAGE) images were acquired parallel with the anterior commissure-posterior commissure (AC-PC) line. The acquisition parameters were TR/TE = 2,200/2.88 ms, flip angle = 13°, matrix = $220 \times 320 \times 208$, and voxel size $0.8 \times 0.8 \times 0.8$ mm.

Each sMRI scan was pre-processed with FSL [17] and ANTs [18]. The processing pipeline includes: (1) bias field correction with N4, (2) linear registration to AAL3 [19] template in the MNI space, resampling to $1 \times 1 \times 1$ mm^3 resolution, and cropping the whole brain to size of $181 \times 217 \times 181$ mm^3, (3) brain extraction, (4) non-linear registration to AAL3 template, and (5) partition ROIs of AAL3 template into the registered sMRI volumes.

2.2 Proposed Method

Subject Grouping. To identify LLD individuals' cognitive change within 5 years, we selected 308 subjects with sMRI scans from the two studies. Using MRI scan time as baseline, category labels were given based on 5-year diagnosis. These subjects were categorized into 4 groups: (1) 88 non-depressed cognitively normal (CN-N) subjects: CN-N subjects at baseline who did not progress to cognitive impairment within the 5-year follow-up; (2) 180 depressed CN (CN-D) subjects: CN-D subjects at baseline but cognitively normal at diagnosis and within the 5-year follow-up; (3) 23 CI subjects: depressed subjects who developed cognitive impairment but not dementia in 5-year follow-up; (4) 17 AD subjects: depressed participants with diagnosis of dementia at 5 years after baseline.

Predictive Deep Classification Network. Inspired by [13], we designed a Predictive Deep Classification Network (PDCN) using Resnet as backbone for prediction of LLD progression to AD. The PDCN model (with Resnet18 as backbone) was shown in Fig. 1, while other networks such as Resnet34 and Resnet50 could also be used as backbones in the proposed PDCN. Accordingly, we had three models, *i.e.*, PDCN18, PDCN34, PDCN50 with Resnet18,

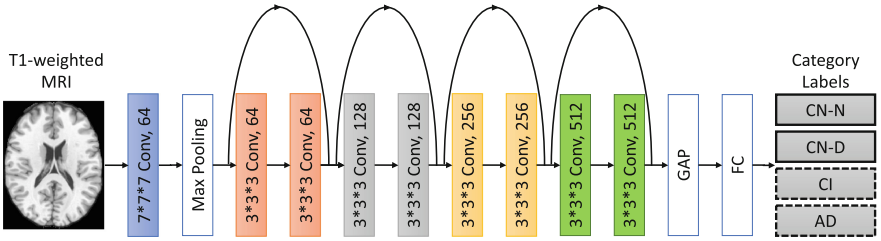


Fig. 1. Architecture of the proposed Predictive Deep Classification Network (PDCN18) for multi-category classification based on T1-weighted MRI. Conv: Convolution; GAP: Global average pooling; FC: Fully connected layer.

Resnet34, Resnet50 as backbones, respectively. The PDCN takes the whole-brain sMRI (size: $181 \times 217 \times 181$) as input and a specific one-hot coding vector as output. The PDCN18 consists of one 3D convolutional (Conv) layer, one max pooling layer, four residual blocks, one global average pooling (GAP) layer, and one fully connected (FC) layer. Compared with PDCN18, the PDCN34 and PDCN50 have similar architectures but more residual blocks [20].

In PDCN, the first 3D Conv layer uses the kernel size of 7, stride of 2, zero padding followed by batch normalization (BN) and rectified linear unit (ReLU) activation. The channel number is 64. The max pooling (MP) layer uses kernel size 3, stride 2 and zero padding. The output of the MP layer is then forwarded to four stacked residual blocks. The residual blocks contain two Conv layers with the kernel size of 3, stride of 1 and zero padding. The BN and ReLU activation are applied to the first Conv layer in the residual block, while only BN is applied to the second Conv layer. A skip connection is applied to sum the input and output of each residual block. The ReLU activation is then applied to the summed output. The channel numbers of the four residual blocks in PDCN18 are 64, 128, 256, and 512, respectively. On the feature maps yielded by the last residual block, a GAP operation is applied to generate a feature vector, which is then fed to the last FC layer to get a probability score vector that indicates the probability of each input sMRI belonging to a specific category.

To handle the small-sample-size issue, we used pre-trained 3D Resnet models [21] to fine tune the proposed PDCN model. The PDCN model was trained with Adam algorithm (batch size: 4, learning rate: 0.0001). All the experiments were implemented with PyTorch under the environment of Python 3.7 on the Ubuntu 18.04 system with NVIDIA TITAN Xp GPU.

3 Experiment

Experimental Setting. Two groups of experiments were performed in this work, including (1) CN-D vs. CN-N classification, and (2) AD vs. CI vs. CN-D vs. CN-N classification. Before model training, all sMRI scans were pre-processed with HistogramStandardization and ZNormalization using TorchIO [22] to avoid

Table 1. Results (in terms of mean±standard deviation) of three different methods in the binary classification task (*i.e.*, CN-D vs. CN-N classification) with T1-weighted MRIs. The term "*-S" denotes a model trained from scratch (*i.e.*, without pretraining).

Method	ACC (%)	SEN (%)	SPE (%)	F1s (%)	AUC (%)
SVM	62.32 ± 3.20	65.55 ± 9.54	55.75 ± 14.86	69.75 ± 4.61	60.65 ± 4.01
RF	63.03 ± 5.98	64.44 ± 12.63	$\mathbf{60.19 \pm 10.81}$	69.54 ± 7.44	$\mathbf{62.32 \pm 3.85}$
PDCN34-S	62.34 ± 6.37	67.77 ± 8.91	51.11 ± 11.73	70.54 ± 6.10	59.44 ± 6.45
PDCN34	$\mathbf{64.98 \pm 8.31}$	$\mathbf{70.55 \pm 14.24}$	53.26 ± 6.10	$\mathbf{72.44 \pm 8.50}$	61.91 ± 5.22

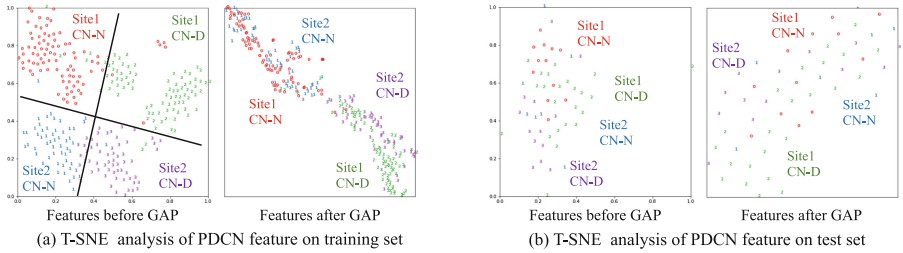


Fig. 2. Features distributions of the last two layers (before and after the GAP layer) in binary classification PDCN using t-SNE on the training set (a) and test set (b). Different colors denote different category labels and imaging sites.

the effect of multi-site data heterogeneity. To balance the data, sMRI scans of CN-N, CI and AD were duplicated and augmented using random affine transform in TorchIO. Five metrics were used to evaluate the prediction performance, including accuracy (ACC), sensitivity (SEN), specificity (SPE), F1 scores (F1s), and the area under the ROC curve (AUC). A 5-fold cross-validation strategy was used in the experiments. Specifically, we first randomly splited all subjects into 5 groups for each category. Then, one group was alternatively used as test set and the remaining 4 groups as training set. The experiments were run 5 times independently to avoid any bias introduced by the random split. The mean and standard deviation results were recorded.

Competing Methods. We compared our PDCN with two popular machine learning methods. (1) Support Vector Machine (**SVM**): SVMs are supervised learning models with associated learning algorithms that analyze data for classification analysis. SVM maps training samples to points in space so as to maximise the width of the gap between the two categories. New samples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. (2) Random Forest (**RF**): It is an ensemble learning method for classification tasks that operates by constructing a multitude of decision trees at training time. The output of the random forest is the category selected by most trees. In both SVM and RF methods, we used the region-of-interest based

Table 2. Performance (in terms of mean ± standard deviation) comparison of our PDCN with different Resnets as backbones in the task of CN-D vs. CN-N classification. The term "*-S" denotes a model trained from scratch (*i.e.*, without pretraining).

Method	ACC (%)	SEN (%)	SPE (%)	F1s (%)	AUC (%)
PDCN18	61.20 ± 4.88	68.88 ± 3.62	45.35 ± 14.31	70.48 ± 3.17	57.12 ± 6.97
PDCN18-S	62.34 ± 5.39	67.77 ± 8.69	51.11 ± 19.29	70.61 ± 4.54	59.44 ± 7.69
PDCN34	**64.98 ± 8.31**	**70.55 ± 14.24**	**53.26 ± 6.10**	**72.44 ± 8.50**	**61.91 ± 5.22**
PDCN34-S	62.34 ± 6.37	67.77 ± 8.91	51.11 ± 11.73	70.54 ± 6.10	59.44 ± 6.45
PDCN50	62.71 ± 8.03	**70.55 ± 9.93**	46.60 ± 6.01	71.52 ± 7.41	58.57 ± 7.09
PDCN50-S	57.49 ± 5.76	61.66 ± 12.93	48.95 ± 13.78	65.44 ± 8.75	55.31 ± 4.28

intensity features to represent each sMRI, yielding a 166-dimensional feature vector for each subject based on the AAL3 template.

Table 3. Performance comparison of PDCN (with the Resnet18 as backbone) in three different classification tasks.

Task	ACC (%)	SEN (%)	SPE (%)	F1s (%)	AUC (%)
CN-D vs. CN-N	**61.20 ± 4.88**	**68.88 ± 3.62**	45.35 ± 14.31	**70.48 ± 3.17**	57.12 ± 6.97
CI vs. CN-D vs. CN-N	52.60 ± 5.34	42.52 ± 3.92	71.66 ± 3.30	42.12 ± 3.67	**63.04 ± 2.84**
AD vs. CI vs. CN-D vs. CN-N	34.62 ± 3.54	28.12 ± 1.84	**75.65 ± 0.73**	20.56 ± 1.67	50.24 ± 1.13

Results of Binary Classification. Table 1 reported the results of different methods in the binary classification task (*i.e.*, CN-D vs. CN-N classification). It could be seen from Table 1 that the RF and our PDCN34 produced the overall better results. For instance, the PDCN34 achieved very good results on (ACC=64.98%; SEN=70.55%; F1s=72.44%). The RF method also achieved good results with SPE=60.19% and AUC=62.32%. The fine tuned PDCN34 could get better overall accuracy while the RF method performed better on the CN-N category what had less training samples. These results implied that deep learning methods were not always better than machine learning methods, especially when the number of samples was limited and the data was unbalanced. By comparing the performance of fine-tuned PDCN34 and the PDCN34 model trained from scratch, we could see that the model pre-trained on multiple large data sets had better feature extraction ability and generalization performance. The above experimental results and analysis proved that it was possible to predict the progression of depression after 5 years based on sMRI features.

Results of Multi-class Classification. To understand the association between LLD and AD, we conducted 3-category classification (*i.e.*, CI vs. CN-D vs. CN-N) and 4-category classification (*i.e.*, AD vs. CI vs. CN-D vs. CN-N) experiments, with results reported in Table 3. In the 3-category classification experiment, CI and AD were combined into CI group. As shown in Table 3, with

the increase of categories, the overall performance of the classification model dropped sharply. In order to further analyze the reasons, we listed the classification indicators of each category, as shown in Table 4. Table 4 showed that the recognition rate of newly added categories was very low, which greatly affected the overall classification accuracy of the model. A reasonable explanation was that the number of samples was too small and even fewer for test.

Table 4. Classification results of PDCN (with Resnet18 as backbone) for each category in one of 5 folds. The term "Support" denotes number of samples for one fold test.

Task	Category	SEN (%)	SPE (%)	F1s (%)	Support
CN-D vs. CN-N	CN-D	66.66	50.00	69.56	36
	CN-N	50.00	66.66	46.15	18
CI vs. CN-D vs. CN-N	CI	12.50	94.44	16.66	**8**
	CN-D	66.66	42.30	64.00	36
	CN-N	38.88	72.72	37.83	18
AD vs. CI vs. CN-D vs. CN-N	AD	0.00	100.00	0.00	**5**
	CI	0.00	100.00	0.00	**4**
	CN-D	72.22	25.86	49.52	36
	CN-N	44.44	77.63	37.20	18

4 Discussion

MRI Feature Distribution Analysis. To interpret results in Table 1 and understand the MRI features extracted by PDCN, the t-SNE method [23] was used to analyze the distribution of intermediate features generated by our PDCN18 in CN-D vs. CN-N classification. The feature maps of last two layers in the network (features before and after GAP layer) were extracted for analysis. The analysis results of t-SNE was shown in Fig. 2. It could be observed from Fig. 2 (a) that the features were separated well by category labels and sites for training data. But the features of the two categories on the test set in Fig. 2 (b) overlapped with each other to some extent. It implied that the PDCN might have over-fitting problems during training. These results further confirmed the impact of limited sample size and data imbalance on deep learning models. Besides, the feature differences between sites were still significant after simply combining multi-site MRI data, thus negatively affecting the generalization ability of PDCN.

Influence of Different Backbones. Table 2 reported the performance comparison results of PDCN in CN-D vs. CN-N classification with different Resnets as backbones. It could be seen from Table 2 that the fine tuned PDCN34 produced the overall best results and the fine tuned PDCN50 performed the second

best. This implied that the depth and complexity of the network structure needed to match the number of samples in the data set to achieve the best classification accuracy and generalization ability. By comparing the performance of fine tuned PDCN34 and the PDCN34 model trained from scratch, we could observe that the fine tuned models always had better performance regardless of the depth of the network. Besides, the input data of PDCN50 was downsampled by $2 \times 2 \times 2$, while there was no significant performance degradation. This may indicated that the input size might not the critical factor for the performance of our method in CN-D vs. CN-N classification.

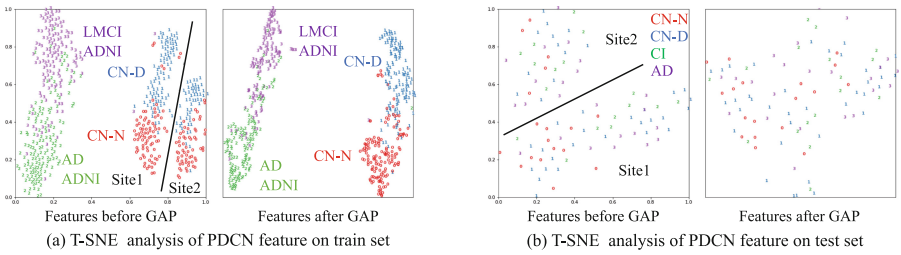


Fig. 3. Features distributions of the last 2 layers in 4-category classification PDCN trained with ADNI data using t-SNE on the train set (a) and test set (b). Different colors denote different categories and imaging sites.

Influence of Data Augmentation. To solve the small-sample-size problem for CI and AD groups, we tried to select 180 subjects with diagnosis of late mild cognitive impairment (LMCI) and 180 subjects with AD from ADNI database for training, and the original CI and AD groups were all used in the test set. Unfortunately, the trained model still couldn't recgonize the CI and AD subjects when testing. We also used t-SNE to analyze the learned MRI features, as illustrated in Fig. 3. It could be observed from Fig. 3 (a) that the features were separated well by category labels and sites when training, but the distribution differences between ADNI and the NCODE/NBOLD was significant. As shown in Fig. 3 (b), the MRI features on test set overlapped to some extent. It implied the PDCN trained with ADNI data couldn't generalize well to new CI and AD data from NCODE and NBOLD, possible due to the significant cross-dataset feature distribution differences. These results suggested that simply augmenting the training sample size by borrowing data from related studies (such as ADNI) might not necessarily improve the discriminative capability of predictive models.

5 Conclusion and Future Work

In this paper, we investigated the cognitive progression of LLD for the first time with deep learning and machine learning methods based on T1-weighted MRI.

Experimental results on 308 subjects with sMRIs acquired from two sites suggested the potential of our method in automated prediction of LLD progression to AD. We further analyzed and discussed the factors that limit the deep learning classifier in the experiments. In the future, we will investigate more suitable feature extraction methods and reasonable learning strategies (*e.g.*, meta learning [24] and domain adaptation [25]) to address the issues of small-sample-size, imbalanced data and cross-site data heterogeneity. Furthermore, we will fuse both functional and structural imaging modalities to discover potential biomarkers for LLD progression prediction and AD detection.

Acknowledgement. This work was supported in part by NIH grant (Nos. AG073297, MH108560, and AG041721). L. Zhang was supported in part by Shandong Provincial Natural Science Foundation (No. ZR2020QF011). This work was finished when Lintao Zhang was visiting the University of North Carolina at Chapel Hill.

References

1. Burke, A.D., Goldfarb, D., Bollam, P., Khokher, S.: Diagnosing and treating depression in patients with Alzheimer's disease. Neurol. Therapy **8**(2), 325–350 (2019)
2. Rashidi-Ranjbar, N., et al.: Frontal-executive and corticolimbic structural brain circuitry in older people with remitted depression, mild cognitive impairment, Alzheimer's dementia, and normal cognition. Neuropsychopharmacology **45**(9), 1567–1578 (2020)
3. Lebedeva, A.K., et al.: MRI-based classification models in prediction of mild cognitive impairment and dementia in late-life depression. Front. Aging Neurosci. **13**, 1–9 (2017)
4. Butters, M.A., et al.: Pathways linking late-life depression to persistent cognitive impairment and dementia. Dialogues in Clinical Neuroscience (2022)
5. Joko, T., et al.: Patterns of hippocampal atrophy differ among Alzheimer's disease, amnestic mild cognitive impairment, and late-life depression. Psychogeriatrics **16**(6), 355–361 (2016)
6. Alexopoulos, G.S.: Mechanisms and treatment of late-life depression. Transl. Psych. **9**(1), 1–16 (2019)
7. Geerlings, M.I., Gerritsen, L.: Late-life depression, hippocampal volumes, and hypothalamic-pituitary-adrenal axis regulation: a systematic review and meta-analysis. Biol. Psychiat. **82**(5), 339–350 (2017)
8. Zhou, T., Thung, K.H., Liu, M., Shen, D.: Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model. IEEE Trans. Biomed. Eng. **66**(1), 165–175 (2018)
9. Hermida, A.P., McDonald, W.M., Steenland, K., Levey, A.: The association between late-life depression, mild cognitive impairment and dementia: Is inflammation the missing link? Expert Rev. Neurother. **12**(11), 1339–1350 (2012)
10. Teodorczuk, A., et al.: Relationship between baseline white-matter changes and development of late-life depressive symptoms: 3-year results from the LADIS study. Psychol. Med. **40**(4), 603–610 (2010)

11. Guan, H., Liu, Y., Xiao, S., Yue, L., Liu, M.: Cost-sensitive meta-learning for progress prediction of subjective cognitive decline with brain structural MRI. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 248–258. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_24
12. Mousavian, M., Chen, J., Greening, S.: Depression detection using feature extraction and deep learning from sMRI images. In: 18th IEEE International Conference On Machine Learning And Applications (ICMLA). IEEE 2019, pp. 1731–1736 (2019)
13. Yang, E., Wang, L., Steffens, D., Potter, G., Liu, M.: Deep factor regression for computer-aided analysis of major depressive disorders with structural MRI data. In: IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE 2021, pp. 208–211 (2021)
14. Jack Jr, C.R., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Resonan. Imaging Official J. Int. Soc. Magn. Resonan. Med. **27**(4), 685–691 (2008)
15. Steffens, D.C., et al.: Methodology and preliminary results from the neurocognitive outcomes of depression in the elderly study. J. Geriatr. Psychiatr. Neurol. **17**(4), 202–211 (2004)
16. Steffens, D.C., Wang, L., Manning, K.J., Pearlson, G.D.: Negative affectivity, aging, and depression: Results from the Neurobiology of Late-Life Depression (NBOLD) study. Am. J. Geriatr. Psychiatr. **25**(10), 1135–1149 (2017)
17. Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M.: FSL. Neuroimage **62**(2), 782–790 (2012)
18. Avants, B.B., Tustison, N., Song, G., et al.: Advanced normalization tools (ANTS). Insight J. **2**(365), 1–35 (2009)
19. Rolls, E.T., Huang, C.C., Lin, C.P., Feng, J., Joliot, M.: Automated anatomical labelling atlas 3. Neuroimage **206**, 116189 (2020)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
21. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
22. Pérez-García, F., Sparks, R., Ourselin, S.: Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput. Methods Programs Biomed. **208**, 106236 (2021)
23. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11), 1–27 (2008)
24. Vilalta, R., Drissi, Y.: A perspective view and survey of meta-learning. Artif. Intell. Rev. **18**(2), 77–95 (2002)
25. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. **69**(3), 1173–1185 (2021)



ClinicalRadioBERT: Knowledge-Infused Few Shot Learning for Clinical Notes Named Entity Recognition

Saed Rezayi[1], Haixing Dai[1], Zhengliang Liu[1], Zihao Wu[1], Akarsh Hebbar[1], Andrew H. Burns[1], Lin Zhao[1], Dajiang Zhu[2], Quanzheng Li[3], Wei Liu[4], Sheng Li[5], Tianming Liu[1], and Xiang Li[3(✉)]

[1] School of Computing, The University of Georgia, Athens, GA, USA
[2] Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX, USA
[3] Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
xli60@mgh.harvard.edu
[4] Department of Radiation Oncology, Mayo Clinic, Phoenix, AZ, USA
[5] Department of Computer Science, University of Virginia, Charlottesville, VA, USA

Abstract. Transformer based language models such as BERT have been widely applied to many domains through model pretraining and fine tuning. However, in low-resource scenarios such as clinical cases, customizing a BERT-based language model is still a challenging task. In this paper, we focus on the radiotherapy domain and train a ClinicalRadioBERT model for analyzing clinical notes through a two-step procedure. First, we fine tune a BioBERT model by exploiting full texts of radiotherapy literature and name this model as RadioBERT. Second, we propose a knowledge-infused few-shot learning (KI-FSL) approach that leverages domain knowledge and trains the ClinicalRadioBERT model for understanding radiotherapy clinical notes. We evaluate ClinicalRadioBERT on a newly collected clinical notes dataset and demonstrate its superiority over baselines on few-shot named entity recognition. We will apply the ClinicalRadioBERT to link BERT and medical imaging for radiotherapy.

Keywords: Biomedical language models · Natural language processing · Named entity recognition

1 Introduction

Transformer-based language models (LM) such as BERT [4] have revolutionized many NLP tasks and their applications in many domains. In general, these NLP models, including biomedical LMs such as BioBERT [13], have been trained in resource-rich environments with large-scale training datasets. Therefore, they may have limited capability in capturing domain knowledge in a specific area with limited resources, such as radiotherapy and radiology clinical notes [17,21,22].

S. Rezayi, H. Dai, Z. Liu and Z. Wu—Co-first authors.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 269–278, 2022.
https://doi.org/10.1007/978-3-031-21014-3_28

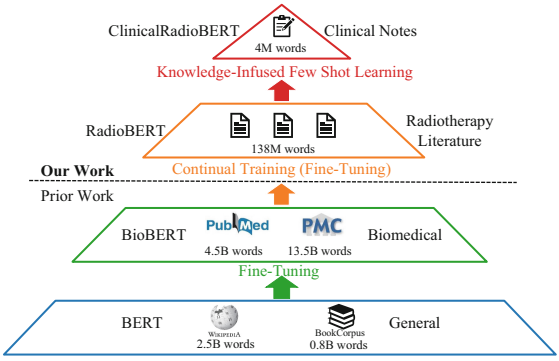


Fig. 1. Our proposed pipeline. Existing works are built on top of BERT and useful for generic biomedical applications. While RadioBERT is specifically designed for radiotherapy applications. Moreover, ClinicalRadioBERT is designed to tackle the issue of low-resource setting in biomedical Named Entity Recognition task using external knowledge.

Table 1. Comparisons of RadioBERT, ClinicalRadioBERT and related BERT models.

Model	Vocabulary	Pretraining	Corpus	Text Size
BioBERT [13]	Wiki+Books	Continual Pretraining	PubMed	4.5B words
SciBERT [2]	PMC + CS	From Scratch	PMC + CS	3.2B words
BlueBERT [18]	Wiki+Books	Continual Pretraining	PubMed + MIMIC	4.5B words
PubMedBERT [7]	PubMed	From Scratch	PubMed	3.1B words
ClinicalBERT [1]	Wiki+Books	Continual Pretraining	MIMIC	0.5B words
RadioBERT	Literature	Continual Pretraining	Literature	138M words
ClinicalRadioBERT	Clinical Notes	-	Clinical Notes	4M words

Recently, few-shot learning has demonstrated great promise in low-resource scenarios by using only a few annotated training samples [6,8,9,20,23,26]. Inspired by these successes, in this work, we focus on the radiotherapy domain and aim to train a ClinicalRadioBERT model for analyzing radiotherapy clinical notes. We partner with clinical practitioners because there is a strong demand for practical NLP solutions in radiation oncology, and in clinical practices in general.

Our main contributions are summarized as follows.

– We propose a novel knowledge-infused few-shot learning (KI-FSL) framework and train a ClinicalRadioBERT model for low-resource NLP scenarios. Our experiments demonstrate that infusing domain expertise into few-shot learning can significantly improve clinical notes named entity recognition. We plan to release our datasets and language models to the community.

- We collected, cleaned, and preprocessed a large-scale corpus of radiotherapy literature with 138 million words. This domain corpus has been instrumental to continually pre-train the BioBERT to RadioBERT. Our RadioBERT model obtains comparable results to state-of-the-art language models on benchmark datasets.
- We collected, anonymized and preprocessed a collection of clinical notes generated by physicians in clinical radiotherapy. This dataset is annotated by domain experts, and our study of this dataset is a real-world clinical scenario of development and evaluation of few-shot learning NLP methods.
- To the best of our knowledge, our ClinicalRadioBERT model is the first large language model specifically designed for radiotherapy. The superior performance of our model can facilitate various downstream NLP applications in this field, and spur future innovations such as linking NLP with imaging.

2 Related Work

2.1 Biomedical Language Models

The adaptation of BERT [4] in the biomedical domain has resulted in extensions such as BioBERT [13], BlueBERT [18], ClinicalBERT [10], PubMedBERT [7] and SciBERT [2]. These BERT-based models are typically developed through continual pretraining [10,13,18], with some being trained from scratch on domain-specific corpora [2,7]. Generally, BERT-based models pre-trained on domain-specific datasets outperform the original BERT [4] on domain-specific tasks [7]. Our experts note that all these applications have practical significance in real-world clinical practices. Specifically, biomedical language models can save time locating, retrieving and organizing important information from clinical text. In addition, they also expect that such language models can help reduce human errors and biases.

2.2 Knowledge Infused Data Augmentation for Language Models

Training data augmentation is a common and intuitive few-shot learning approach [23]. However, there are conflicting results regarding the effectiveness of applying data augmentation to pre-trained language models. For example, EDA [24], a popular and intuitive data augmentation framework with word-level replacements offers limited performance gains for BERT and RoBERTa [5]. On the other hand, contradictory reports exist. A study done by Shi et al. [21] shows that simply interpolating phrases from training data sentences significantly improves the performance of pre-trained models. In general, NLP researchers advocate for more research on data augmentation with large pre-trained models [5].

In this study, we find that our knowledge-infused few-shot learning approach significantly improves BERT-based models' NER performance with real-world clinical data. Our results reveal a potentially promising application scenario for

Table 2. Statistics of text corpus for RadioBERT.

Journal	#papers	#docs	#words
Acta Oncologia	5,846	2.1M	40M
British Journal of Cancer	306	95K	2.6M
radiation oncology journal	2,810	1.2M	24M
Journal of Clinical Oncology	522	149K	4.1M
Journal of Hematology and Oncology	1,103	514K	9.7M
Lancet	1,110	111K	3.8M
Nature Reviews Clinical Oncology	195	6.5K	114K
Radiation Oncology	775	333K	8.7M
Radiologica Medica	1,873	468K	13M
Strahlentherapy and Onkologie	2,985	682K	13M
Technology in Cancer and Research	1,886	750K	14M
Technology in Cancer Research and Treatment	388	189K	3.9M
Total	19,799	6.6M	138M

augmenting pre-trained language models with a knowledge graph. We believe this work **contributes to data augmentation research** in addition to its significance in clinical NLP.

3 RadioBERT

There is a huge range of text corpora available in the biomedical and clinical domains. Table 1 summarizes some representative BERT models for biomedical applications as well as the corresponding corpora, such as PubMed and MIMIC. However, these text corpora are not suitable for our task of analyzing radiotherapy clinical notes for two reasons: First, they mainly use abstracts of publications and thus ignore comprehensive information in full texts. Second, these text corpora cover a wide range of biomedical topics but are not tailored to radiotherapy.

These observations motivate us to collect a new text corpus focusing on radiotherapy. In particular, we considered various Oncology and Radiation Oncology journals in creating our dataset. We performed a comprehensive web scraping of these journals to build our dataset. Table 2 shows the statistics of our text corpus. We adopt a pre-trained BioBERT model and use a masked language modeling strategy to further fine-tune it using our dataset. For efficiency purposes and to fine-tune the model, we unfreeze the last two layers of the BioBERT model and freeze the rest of the layers.

4 ClinicalRadioBERT

4.1 Clinical Notes Dataset

Data Collection. The raw corpus we used was from the clinical notes generated by physicians in a department of radiation therapy of a large medical center. In

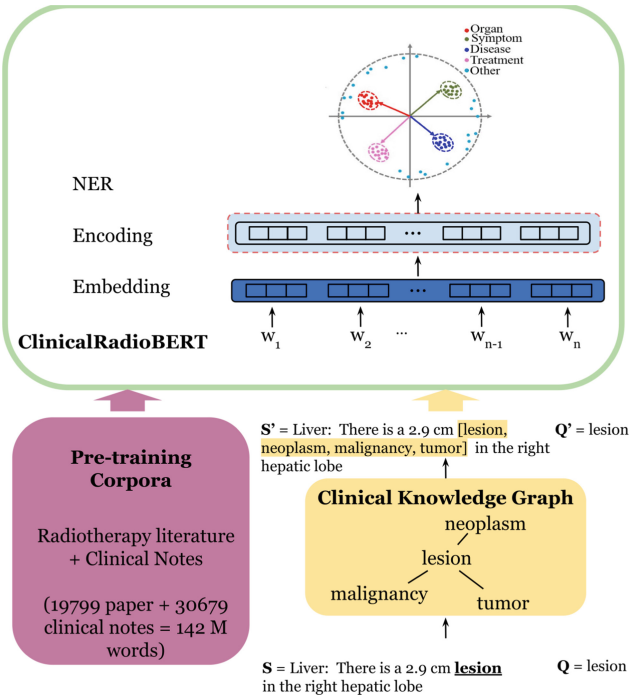


Fig. 2. The overall framework of ClinicalRadioBERT, which is trained on radiotherapy literature and clinical notes from the collaborating research hospital. We evaluate ClinicalRadioBERT on the Named Entity Recognition (NER) task. The NER component has two inputs: 1. The sentences from the processed texts 2. New entities obtained by querying our knowledge graph with text from the clinical notes. The output of the framework is a label that assigns a token to a class of entities. In this figure, "S" denotes a sentence, and "Q" denotes a query term for the knowledge graph

total, we collected 30,678 clinical anonymized notes under IRB approval. The raw data contain approximately 4 million words and are unlabeled, which are used to pre-train our model. Given the limited capacity of our experts, instead of performing a comprehensive entity distribution analysis among all of these clinical notes, we randomly selected 60 clinical reports with 10,433 words and 13,431 tokens for few-shot learning experiments in this work.

Expert Annotation. Named entities in radiotherapy are quite context-dependent, and thus annotation of these named entities is complicated. In this work, 3 experts with substantial knowledge in both radiotherapy and NLP manually annotated randomly selected samples. For each token, the final labeling decision is reached only after reconciliation, which provides a strong guarantee for inter-annotator agreement.

We focus on four categories of named entities that are clinically relevant, i.e., human organs (labeled as B-O), diseases (labeled as B-D), symptoms (labeled as B-S), and treatments (labeled as B-T). Any other entities will be labeled as (O). Some examples of these annotations are: "enlarged" (B-S), "prostate"

(I-S); "urinary" (B-S), "tract" (I-S), "symptom" (I-S); "kidney" (B-O); "cholecystectomy" (B-T); "additional" (O). In total, we manually annotated 6,300 tokens (from the 60 selected clinical reports) that were then used for domain knowledge-based data augmentation.

4.2 Data Augmentation

We use entity linking techniques to introduce new and related entities to the existing annotated named entities. By incorporating new knowledge into the data, we extend the semantic space of the existing entities. For instance "pancreatic lesions" is a named entity in our dataset that was further linked to "pancreatic cancer" and "pancreatic tumor". To add new entities we first identify the sequence of tokens that form a named entity and use that as a query for the entity linking task. We find 10 new entities per original named entity and include them in the original dataset. For this experiment, we use BLINK [25] to build our clinical knowledge graph, but any other entity linking algorithms would work [12].

4.3 KI-FSL

The major advantage of few-shot learning in NLP is that high cost related to data labeling is eliminated, which is quite important in a biomedical NLP scenario and motivated us to develop the improved few-shot learning method KI-FSL in this work.

We adopted the recently developed StructShot few-shot learning method in [26] as our NER method. StructShot extends the nearest neighbor classification with contextualized representations of tokens in the labelled support set. Our method improves the StructShot method in two aspects. First, our domain-specific RadioBERT model trained on radiotherapy literature corpus is used as the underlying architecture, and randomly selected clinical notes examples are manually labeled as the support set by experts. Second, we infuse domain knowledge into the manually annotated support set by extensively augmenting them. Our experimental results will prove that this data augmentation step of domain knowledge infusion for the support set strategy works very well.

In addition, it should be noted that our language model and the analysis pipeline is compatible with any machine learning-based NER methods. It is not limited to the nearest-neighbor approach adopted by StructShot. This is because the power of our model lies within the multi-step continual pre-training pipeline, which leads to deep semantic understanding and knowledge incorporation of in-domain data.

5 Experiments

In this section we first evaluate our trained language model on 8 benchmark datasets and then in Sect. 5.2 we show the practical value of ClinicalRadioBERT in addressing the task of few-shot named entity recognition.

Table 3. Results of NER on 8 benchmark dataset using RadioBERT and three related models. Best results are marked as bold, and second best results are underlined.

Datasets	RoBERTa			XLM-RoBERTa			RadioBERT			BioBERT-reported			BioBERT-local		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
NCBI-disease	84.65	85.62	85.13	76.10	80.94	78.45	85.84	89.06	**87.42**	88.22	91.25	89.71	84.26	88.12	86.15
LINNAEUS	89.49	80.25	84.62	82.88	67.90	74.65	88.59	81.30	84.79	90.77	85.83	88.24	92.17	84.65	**88.25**
JNLPBA	71.03	80.44	75.44	66.19	78.02	71.62	70.01	82.40	75.70	72.24	83.56	77.49	69.72	83.15	**75.85**
BC5CDR-disease	79.17	81.89	80.51	76.00	78.39	77.18	82.47	84.97	83.70	86.47	87.84	87.15	83.67	87.12	**85.36**
BC5CDR-chem	90.36	88.43	89.38	86.20	85.22	85.95	92.72	92.84	92.60	93.68	93.26	93.47	92.73	93.57	**93.15**
BC4CHEMD	89.51	86.25	87.85	87.30	84.90	86.08	90.94	89.01	89.96	92.80	91.92	92.36	91.42	91.34	**91.38**
BC2GM	80.18	80.25	80.22	72.71	76.54	74.57	81.99	82.28	**82.13**	84.32	85.12	84.72	80.42	82.74	81.56
Species-800	68.81	73.92	71.27	51.73	58.41	54.87	65.27	69.10	67.13	72.80	75.36	74.06	67.38	77.84	**72.23**

5.1 Named Entity Recognition

Table 3 presents the results of RadioBERT and three other baselines, including RoBERTa [15], XLM-RoBERTa [3] and BioBERT [13], on eight biomedical NER benchmark datasets. We report the performance in terms of precision, recall and F1-score. As shown in Table 3, the performance of RadioBERT is on par with state-of-the-art biomedical language models, although it is trained on very specific literature in cancer radiotherapy.

5.2 Few Shot Named Entity Recognition

In addition, Table 4 summarizes our experimental results on few-shot named entity recognition. Specially, we evaluated multiple biomedical BERT models, with or without few shot learning support set augmentation (either small or large scale augmentation, with 200 and 500 annotated entities, respectively). The results in Table 4 clearly indicate that the results obtained with RadioBERT as the baseline model are substantially better than those with BioBERT as the baseline model, demonstrating the significance of fine-tuning with our radiotherapy literature corpus and clinical dataset. More importantly, the results in Table 4 also show that the ClinicalRadioBERT model with large support data augmentation is significantly superior to the others, confirming that our proposed KI-FSL method is effective in clinical notes NER tasks.

5.3 Discussion

Our results show that continuous training using domain-specific corpus can incrementally improve more general NLP models. Moreover, the combined steps of continuous fine-turning and few-shot learning offer an effective approach to domain-specific NLP applications. Specifically, we identify the following open questions and opportunities that can be potentially addressed by adapting this work to more application scenarios:

Table 4. Few-shot NER results on clinical notes using four language models and different settings.

Method	Precision	Recall	F1
PubMedBERT (StructShot)	0.066	0.051	0.058
PubMedBERT (StructShot with small support set augmentation)	0.083	0.062	0.071
PubMedBERT (StructShot with large support set augmentation)	0.116	0.089	0.101
BioBERT (StructShot)	0.157	0.174	0.165
BioBERT (StructShot with small support set augmentation)	0.167	0.213	0.187
BioBERT (StructShot with large support set augmentation)	0.169	0.233	0.196
ClinicalRadioBERT (StructShot)	0.159	0.222	0.185
ClinicalRadioBERT (StructShot with small support set augmentation)	0.159	0.267	0.199
ClinicalRadioBERT (StructShot with large support set augmentation)	**0.180**	**0.311**	**0.229**

Language Models for Semantics-Rich Diseases. Certain diseases such as Alzheimer's disease (AD) are related to speech disorders and cognitive decline [19]. They are inherently connected to core natural language topics, and their rich semantic space invites research in detecting, diagnosing, and analyzing these diseases [11,16,19]. For example, Liu et al. [16] predicted the occurrence of Alzheimer's disease using transcripts from referential communication tasks (RCT) of possible AD patients. However, there is still limited use of large state-of-the-art language models, such as those based on BERT [19]. Our model paves the way for medical researchers to adopt similar pipelines for developing effective language models viable for on-demand downstream task solutions.

Advanced Language Tasks in Radiation Therapy. There are many NLP tasks in radiation therapy, and the NER task discussed in this paper is only one of the many possible applications. Our model can be easily modified and extended to solve a multitude of language tasks, including but not limited to document classification, sentiment analysis, and relation extraction. Furthermore, it can also be used for advanced domain-specific tasks such as knowledge discovery, treatment planning and report generation [14], which are tasks that require deep semantic understanding of the underlying clinical data.

6 Conclusions

In this paper, we proposed the knowledge-infused few-shot learning (KI-FSL) approach and trained the ClinicalRadioBERT model for radiotherapy clinical notes named entity recognition. Experimental results validated the effectiveness of ClinicalRadioBERT. Our results indicate that continuous training using domain-specific corpus can improve more general NLP models and that few-shot learning is a promising approach to further adapt the general NLP models to more specific domains. Finally, the combined steps of continual fine-tuning

and knowledge-infused few-shot learning offer an effective approach to domain-specific NLP applications. In the future, we plan to further integrate NLP with medical imaging for radiotherapy and radiology.

References

1. Alsentzer, E., et al.: Publicly available clinical bert embeddings. In: Clinical NLP Workshop, pp. 72–78 (2019)
2. Beltagy, I., Lo, K., Cohan, A.: Scibert: a pretrained language model for scientific text. In: EMNLP-IJCNLP, pp. 3615–3620 (2019)
3. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: ACL, pp. 8440–8451 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. In: NAACL, pp. 4171–4186 (2019)
5. Feng, S.Y., et al.: A survey of data augmentation approaches for NLP. In: ACL-IJCNLP, pp. 968–988 (2021)
6. Fritzler, A., Logacheva, V., Kretov, M.: Few-shot classification in named entity recognition task. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, pp. 993–1000 (2019)
7. Gu, Y., et al.: Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthc. (HEALTH) **3**(1), 1–23 (2021)
8. Hou, Y., et al.: Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In: ACL, pp. 1381–1393 (2020)
9. Huang, J., et al.: Few-shot named entity recognition: an empirical baseline study. In: EMNLP, pp. 10408–10423 (2021)
10. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 (2019)
11. Hussein, K.I., et al.: Natural language processing to identify patients with cognitive impairment. medRxiv (2022)
12. Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-end neural entity linking. In: CoNLL, pp. 519–529 (2018)
13. Lee, J., et al.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
14. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6666–6673 (2019)
15. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
16. Liu, Z., Paek, E.J., Yoon, S.O., Casenhiser, D., Zhou, W., Zhao, X.: Detecting Alzheimer's disease using natural language processing of referential communication task transcripts. J. Alzheimer's Disease (Preprint) 1–14 (2022)
17. Liu, Z., et al.: Survey on natural language processing in medical image analysis. J. Cent. S. Univ. Med. Sci. **47**(8), 981–993 (2022)
18. Peng, Y., Yan, S., Lu, Z.: Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. In: BioNLP Workshop, pp. 58–65 (2019)
19. Petti, U., Baker, S., Korhonen, A.: A systematic literature review of automatic Alzheimer's disease detection from speech and language. J. Am. Med. Inform. Assoc. **27**(11), 1784–1797 (2020)

20. Rezayi, S., et al.: Agribert: knowledge-infused agricultural language models for matching food and nutrition. In: IJCAI (2022)
21. Shi, H., Livescu, K., Gimpel, K.: Substructure substitution: structured data augmentation for NLP. In: ACL-IJCNLP, pp. 3494–3508 (2021)
22. Thrall, J.H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K., Brink, J.: Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. J. Am. Coll. Radiol. **15**(3), 504–508 (2018)
23. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: a survey on few-shot learning. ACM Comput. Surv. (CSUR) **53**(3), 1–34 (2020)
24. Wei, J., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: EMNLP-IJCNLP, pp. 6382–6388 (2019)
25. Wu, L., Petroni, F., Josifoski, M., Riedel, S., Zettlemoyer, L.: Scalable zero-shot entity linking with dense entity retrieval. In: EMNLP, pp. 6397–6407 (2020)
26. Yang, Y., Katiyar, A.: Simple few-shot named entity recognition with structured nearest neighbor learning. In: EMNLP, pp. 6365–6375 (2020)



Graph Representation Neural Architecture Search for Optimal Spatial/Temporal Functional Brain Network Decomposition

Haixing Dai[1], Qing Li[2,3], Lin Zhao[1], Liming Pan[4], Cheng Shi[5], Zhengliang Liu[1], Zihao Wu[1], Lu Zhang[6], Shijie Zhao[7], Xia Wu[2,3], Tianming Liu[1], and Dajiang Zhu[6(✉)]

[1] School of Computing, The University of Georgia, Athens, GA, USA
[2] School of Artificial Intelligence, Beijing Normal University, Beijing, China
[3] Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education, Beijing, China
[4] School of Computer and Electronic Information, Nanjing Normal University, Nanjing, China
[5] Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland
[6] Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX, USA
dajiang.zhu@uta.edu
[7] School of Automation, Northwestern Polytechnical University, Xi'an, China

Abstract. Decomposing the spatial/temporal functional brain networks from 4D functional magnetic resonance imaging (fMRI) data has attracted extensive attention. Among all these efforts, deep neural network-based methods have shown significant advantages due to their powerful hierarchical representation ability. However, the network architectures of those deep learning models are manually crafted, which is time consuming and non-optimal. This paper presents a novel graph representation neural architecture search (GR-NAS) method based on graph representation to optimize the vanilla RNN cell structure for decomposing spatial/temporal brain networks. The core idea is to embed the *discrete* search space of the RNN cell into a *continuous* domain that preserves the topological information. After that, popular search algorithms, e.g., reinforcement learning (RL) and Bayesian optimization (BO), can be employed to find the optimal architecture in this continuous space. The proposed method was evaluated on the Human Connectome Project (HCP) task fMRI datasets. Extensive experiments demonstrated the superiority of the proposed model in brain network decomposition both spatially and temporally. To our best knowledge, the proposed model is among the early efforts using NAS strategy to optimally decompose spatial/temporal functional brain networks from fMRI data.

Keywords: Graph Representation Neural Architecture Search · Brain network decomposition · fMRI

H. Dai and Q. Li - Co-first authors.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 279–287, 2022.
https://doi.org/10.1007/978-3-031-21014-3_29

1 Introduction

Exploring functional brain networks (FBNs) has been an active research topic in neuroimaging field for years [1, 2]. To decompose the FBNs from functional magnetic resonance imaging (fMRI) data, various model-driven and data-driven methods have been proposed [3–8], among which deep learning-based approaches have gained increasing attention because of their powerful and hierarchical representation ability. For example, a deep convolutional auto-encoder (DCAE) was proposed to explore both the high-level and low-level FBNs [4]. Another group of studies employed a deep belief network (DBN) to identify the FBNs in a hierarchical manner [5, 6]. Recently, recurrent neural network (RNN) based models have shown significant advantages in modeling temporal dependencies within fMRI data and achieved promising results [7, 8]. For instance, a deep sparse recurrent autoencoder (DSRAE) [8] was developed to simultaneously decompose the FBNs at connectome-scale, demonstrating the effectiveness of RNN models in extracting neuroscientifically meaningful spatial/temporal networks from 4D fMRI data.

However, current deep learning models such as the abovementioned DSRAE are limited in the sense that their network architectures are manually crafted. Generally, designing an appropriate or optimal neural network is a laborious and time-consuming process that greatly depends on rich domain knowledge and experience. Also, for modeling the FBNs under different task stimuli, the neural network architectures might need to be optimized, respectively. For example, the optimal hyper-parameters of DSRAE for the emotion task might not be optimal for the working memory task process. To overcome these challenges, some literature efforts were made based on evolutionary neural architecture search (NAS) to find the optimal architectures for 4D fMRI data [9]. Another literature work that adopted differentiable neural architecture search (DARTS) [10, 11], named ST-DARTS [12], was proposed to improve the efficiency of searching optimal RNN architectures while maintaining comparable performances. Nevertheless, ST-DARTS is still limited in discrete space and neglects the topological information within RNN cells, which might be trapped on a local optimum and thus degenerates the performance.

This paper proposed a novel graph representation NAS (GR-NAS) method to optimize the vanilla RNN cell structure for decomposing the spatial/temporal FBNs. Specifically, we represented the RNN cell architectures in the discrete DARTS search space as graphs and embedded them into a latent continuous search space via a Graph Isomorphism Network (GIN) [13] encoder. With the embeddings in this continues space, several search strategies, such as reinforcement learning (RL) [14] and Bayesian optimization (BO) [15], can be adopted to search for the optimal RNN cells. We evaluated the proposed unsupervised NAS framework on publicly available Human Connectome Project (HCP) task fMRI data with RL and BO search strategies. Extensive experimental results demonstrated the superiority of the proposed method in both searching optimal network architectures and decomposing meaningful FBNs. Moreover, the transfer learning process based on unsupervised NAS between different tasks can also achieve good performance, suggesting the robustness and generality of the proposed method.

2 Materials and Method

2.1 Overview

Figure 1 presents the overview of our proposed GR-NAS model. The general target of this work is to learn the optimal RNN cells for decomposing the spatial/temporal FBNs from fMRI data. RNN cells, which have graph-structured architectures, are embedded with the Graph Isomorphism Autoencoder to learn the pre-trained embeddings. Then, two different downstream architecture search algorithms are adopted to search the optimal architecture on the pre-trained embeddings. Finally, we apply the learned RNN cell to 4D fMRI data to obtain the feature map in the latent layer for decomposing the spatial/temporal brain function network.

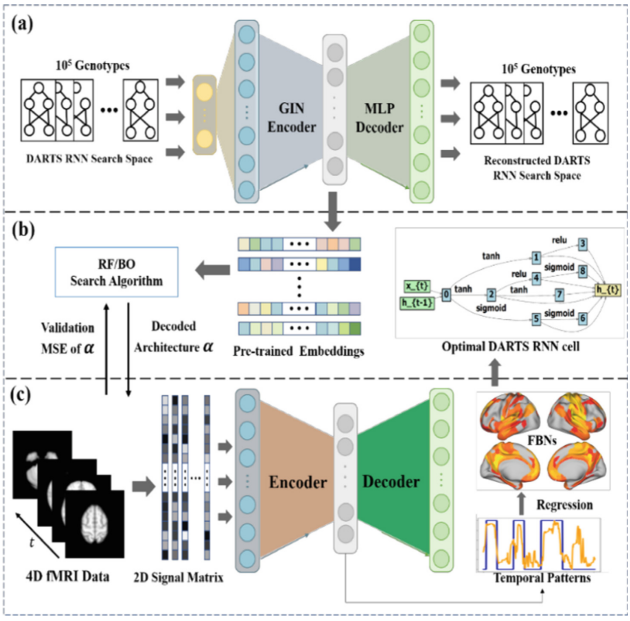


Fig. 1. Our framework of GR-NAS. (a) GIN encoder and multiple layer perceptron (MLP) [17] decoder are used to embed the DARTS RNN cell architectures to obtain pre-trained embeddings. (b) RF/BO search algorithms are used on the pre-trained embeddings to search for the optimal architecture based on the architecture performance estimation. (c) Decomposed spatial/temporal function network with the learned architecture. The input is our 4D-fMRI data which consists of the spatial (360 region of interests) and temporal (length of fMRI tasks' time duration) information. The selected DARTS RNN cell is evaluated by cross-entropy loss.

2.2 Data Description and Pre-processing

We adopted the publicly available HCP grayordinate-based tfMRI datasets from the Q3 release of the Human Connectome Project (https://db.humanconnectome.org). The

detailed acquisition parameters are as follows: TR = 720 ms, TE = 33.1 ms, flip angle = 52°, in-plane FOV = 208 mm × 180 mm, 104 × 90 matrix, slice thickness = 2 mm, 72 slices, multiband factor = 8, echo spacing = 0.58 ms, BW = 2290 Hz/Px. Important preprocessing steps including spatial smoothing, temporal filtering, nuisance regression, and motion censoring were applied to all subjects. In addition, we extracted the 4D fMRI volume of each subject and rearranged it into a 2D signal matrix for Emotion and Working Memory (WM) task, respectively. In the training stage, we randomly selected 750 subjects and divided them into three groups: 450 subjects as training set, 150 subjects as validation set and another 150 subjects as testing set. During the network architecture search process, we only used the training set and validation set. All subjects were included in evaluating the searched architectures.

2.3 Graph Representation Neural Architecture Search

In this work, we proposed a novel GR-NAS model for searching the optimal architecture of RNN cell, and it has been applied in Sec. 2.4 to decompose the FBNs. Specifically, GR-NAS has a variational graph isomorphism autoencoder that embeds the represented graph of RNN cell into a continuous searching space. Then, search algorithms can be used to explore the optimal architecture.

2.3.1 Variational Graph Isomorphism Autoencoder

We set the searching space of our RNN cell as that of DARTS [10]. More details of RNN cells in DARTS search space are referred to [10]. In this space, each RNN cell can be represented as a directed acyclic graph (DAG) $G = (V, E)$, where V is the set of nodes and E is the set of edges. Each node is associated with one of K predefined operations. Therefore, the DAG can be represented by two matrices: one upper triangular adjacency matrix $A \in \mathbb{R}^{N\times N}$ encoding the connections among N nodes and another one-hot operation matrix $X \in \mathbb{R}^{N\times K}$ recording the operation in each node. Then, a two-layer GIN encoder is used to embed the adjacency matrix A and operation matrix X into a vector $Z \in \mathbb{R}^N$. The encoder is defined as:

$$q(Z|X, A) = \prod_{i=1}^{N} q(z_i|X, A), \text{with} q(z_i|X, A) = \mathcal{N}(z_i|\mu_i, \text{diag}(\sigma^2)) \tag{1}$$

where μ, σ are the mean and the variance of approximation $q(z_i|X, A)$, respectively. Then we use a L-layer GIN to obtain the node embedding matrix H:

$$H^{(k)} = MLP^k\left(\left(i + \epsilon^k\right) \bullet H^{(k-1)} + A \bullet H^{(k-1)}\right), k = 1, 2, \ldots, L, \tag{2}$$

where $H^{(0)} = X$, ϵ is a trainable bias, and MLP is a multi-layer perceptron with each layer a linear-batchnorm-ReLU triplet. We use $H^{(L)}$ to obtain the mean μ and the variance σ of the $q(Z|X, A)$. After the encoder, a generative model, which aims at reconstructing $\widehat{X}, \hat{A}$ from the latent variable Z, is defined as:

$$\mathrm{P}(\hat{\mathrm{A}}|Z) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(\hat{\mathrm{A}}_{ij}|z_i, z_j), \text{ with } p(\hat{\mathrm{A}}_{ij} = 1|z_i, z_j) = \sigma\left(z_i^T z_j\right) \tag{3}$$

$$p\left(\hat{X} = [k_1, \ldots, k_n]^T | Z\right) = \prod_{i=1}^{N} P(\hat{X}_i = k_i | z_i) = \prod_{i=1}^{N} softmax(WZ + b) \tag{4}$$

where σ is the logistic sigmoid function and A_{ij} contains the elements of A. With the autoencoder, we maximize the lower bound L:

$$L = E_{q(Z|X,A)}\left[\log p(A|Z)\right] - KL[q(Z|X, A)||p(Z), \tag{5}$$

where we assume that given the latent variable Z, the adjacent matrix A and the operation matrix X are conditionally independent. In other words, $p(X, A|Z) = p(A|Z)p(X|Z)$. The term KL is the KL divergence, which measures the difference between the posterior distribution $q(\bullet)$ and the prior distribution $p(\bullet)$. We perform full-batch gradient descent and use the reparameterization scheme to inject random noise to the training layer, which has been proven to be effective on the regularization of neural networks. The loss function is optimized using mini-batch gradient descent over neural architectures.

2.3.2 Search Strategies

This paper uses two different representative search algorithms to evaluate our model on pre-trained embeddings: reinforcement learning (RL) and Bayesian optimization (BO). In GR-NAS (RL), the embeddings are agents, the validation loss for each embedding is used as the reward, a single layer long short-term memory (LSTM) is used as the controller, and the action is used as the movement on one of the embedding's 16 dimensions. We leveraged the pre-trained embedding pass to the policy LSTM network to evaluate the current state and then obtain the next action and state based on L^2 distance to minimize the reward, i.e., the validation loss of selected embedding. We used the Adam optimizer and set the learning rate as 0.01. The discount factor is set to 0.9, and the baseline value is set to 25. The estimated wall-clock time for each run is set to 20 minutes. In GR-NAS (BO), deep networks are used for global optimization (DNGO) [16] to search the optimal architecture on our embeddings. A one-layer adaptive basis regression network with 128 hidden dimensions is employed for modeling the distribution over function. During the training, selected pre-trained embeddings will receive corresponding loss values and then pass them to DNGO for selecting top architectures.

2.4 Temporal/Spatial Functional Network Learning

After GR-NAS searching on the embeddings, optimal RNN cells are found for spatial/temporal function network learning. The temporal network dynamics were learned with the latent RNN cells and the spatial networks were derived by such temporal network dynamics with Elastic Net regression [18]. Then we put these architectures into our temporal function network learning model, which is the same as the model used for the performance estimate strategy in Fig. 1(c). In the process of architecture searching, we set the epoch as 10.In the temporal function network learning, the epoch was set as 200 to guarantee the convergence with 128 batch size. For other hyperparameters, each cell has 8 nodes, and the number of operations is 4 (tanh, identity, sigmoid, ReLU). The hidden size is set as 32, and the initial learning rate is set as 20. We implemented the proposed GR-NAS model with PyTorch 1.4.1 on a single RTX 2080 GPU.

3 Results

3.1 DARTS Cell Structure and Spatial/Temporal Functional Networks

Compared to ST-DARTS, we implemented GR-NAS to embed and then learn the best cell structure with RL and BO. We can learn the spatial/temporal functional brain networks from WM task and emotion fMRI datasets with the learned cell to learn the cell structure, GR-NAS takes approximately one GPU day. Under different settings in search algorithm parameters and different tasks, the learned cell structures might be distinct. Under the same setting, comparison between our GR-NAS model and ST-DARTS was made on WM fMRI data. As shown in Fig. 2, for different task designs, our GR-NAS model under RL and BO both achieved more remarkable performance than ST-DARTS, which indicates our model can learn the temporal function networks that measured by Pearson correlation coefficient (PCC) better (Fig. 3).

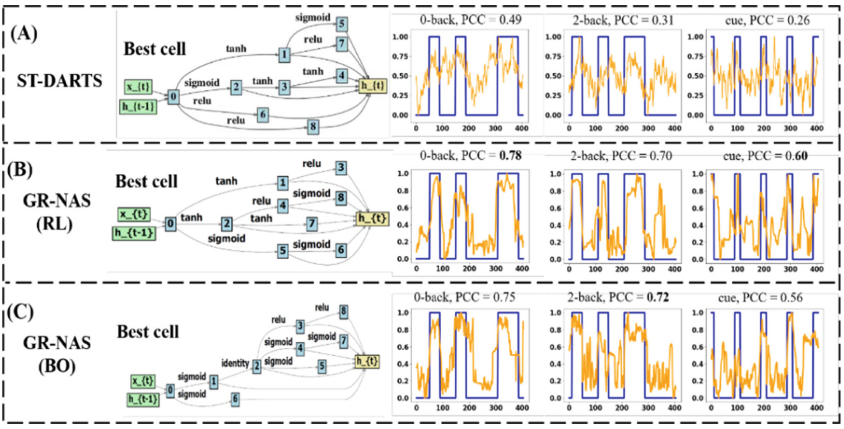


Fig. 2. Comparison pf different cells found by ST-DARTS, GR-NAS (RF), and GR-NAS (BO). The best cells found by various models are placed in the left of each row. Furthermore, the most stimuli-correlated temporal networks learned by given cells on WM different tasks are on the right. Blue curves denote the task-design, and orange curves denote the learned temporal network dynamics (Color figure online).

We use Dice coefficient (DC) [19] to measure the similarity between two sets. DC is commonly used for similarity measuring between the spatial networks and the benchmark, which are derived by Elastic Net regression of fMRI data. Our first preparation is converting the functional spatial networks $Net^{(i)}$ into a Boolean type. Then, we assign one to the activated voxel, whose signal value is more than 10^{-3}, and zero to the deactivate voxel, whose signal value is less than 10^{-3}. Thus, the DC between $Net^{(i)}$ and $Net^{(j)}$ is defined as follows:

$$DC_{i,j} = \frac{2 \times |Net^{(i)} \cap Net^{(j)}|}{|Net^{(i)}| + |Net^{(j)}|} \tag{6}$$

Here we show the results of WM task. As shown in Fig. 4, the individual best is as high as 0.73 in WM, comparing to the highest value 0.58 by ST-DARTS.

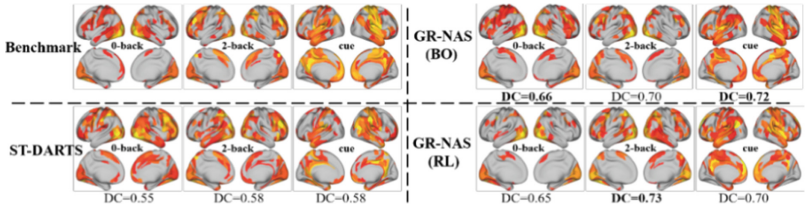


Fig. 3. The most stimuli-correlated spatial networks learned by GR-NAS on WM fMRI tasks. The spatial benchmarks are learned by Elastic Net regression based on the temporal task-design, as in the same way we learned spatial networks. The bottom left contains the result of ST-DARTS.

3.2 Transfer Learning on Different Task-fMRI Datasets

We also tried to use the architecture learned from the emotion fMRI data on WM data to examine the results of decomposing temporal functional brain networks. Figure 4 shows the results of the spatial network learned by transfer learning compared with the ST-DARTS benchmark. Compared with the results in Fig. 2 and Fig. 4, the average correlation rates of transfer learning are slightly lower than the proposed method, but it still outperforms ST-DARTS because of an embedding process that makes the similar characteristics between the two tasks into a same target domain. This result demonstrated that our embedding method can effectively find an optimal architecture.

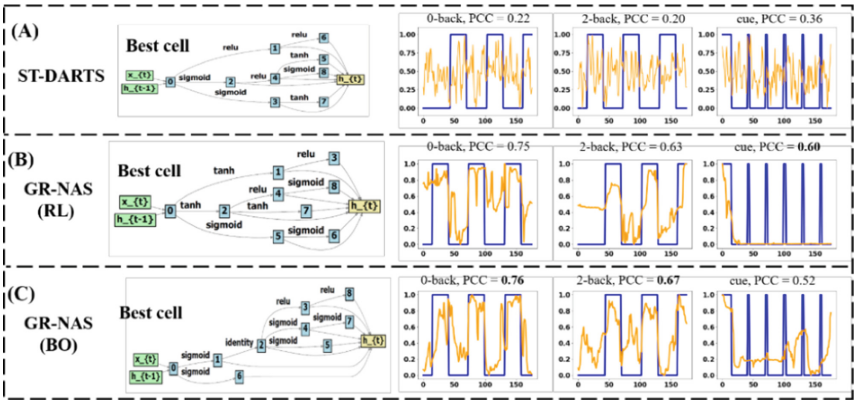


Fig. 4. Comparison between different cells found by ST-DARTS, GR-NAS (RL) and GR-NAS (BO). The best cells found by the various models are placed in each row's left. Moreover, the most stimuli-correlated temporal networks learned by given cells on WM different tasks are on the right. Blue curves denote the task-design, and orange curves denote the learned temporal network dynamics (Color figure online).

3.3 Stability and Robustness of GR-NAS

To illustrate the stability and robustness of the GR-NAS model, we showed all the three runs' results of PCCs under GR-NAS between learned temporal network dynamics and

task design. In Fig. 5, the PCCs of GR-NAS vary from 0.6 to 0.8, which is substantially higher than the previous results by ST-DARTS from 0.2 to 0.4. This result indicates that the proposed GR-NAS model can stably and robustly derive meaningful networks. The DCs of BO NAS and RF NAS are from 0.65 to 0.75 across the WM task, which outperform 0.55 to 0.65 by ST-DARTS. This result indicates that the spatial networks learned from the proposed GR-NAS model can derive similar maps with the benchmark even with different architecture search algorithms. Additionally, the variance of PCC/DC values under the proposed GR-NAS model is much less than that of PCC/DC values by ST-DARTS (Fig. 5), which suggests the effectiveness and robustness of GR-NAS.

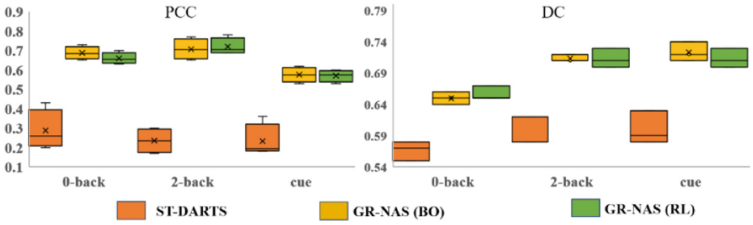


Fig. 5. The temporal network dynamics' performance of ST-DARTS, GR-NAS (BO) and GR-NAS (RL) during four runs. Each box reveals the maximum (upper whisker), minimum (bottom whisker), median (X mark), first quantile (box bottom) and third quantile (box top).

4 Discussion and Conclusion

This paper presented a novel GR-NAS model for optimal brain network decomposition. Unlike previous methods, we embedded the DARTS RNN cells into a pre-trained embedding space to preserve the topological information and learn the optimal architecture on a continuous space. Then, we implemented RL and BO in GR-NAS model to search for the optimal architectures based on the embeddings. In addition, we implemented transfer learning in this work to evaluate our GR-NAS model. It is promising that transfer learning is better than ST-DARTS, but it is slightly worse than the proposed GR-NAS model. The results indicate that our continuous embedding search space is better than a discrete search space, and it can effectively find the optimal architecture for deep neural networks.

References

1. Woolrich, M.W., Jenkinson, M., Brady, J.M., Smith, S.M.: Fully Bayesian spatio-temporal modeling of fMRI data. IEEE Trans. Med. Imaging **23**(2), 213–231 (2004)
2. Zhao, Y., Dai, H., Zhang, W., Ge, F., Liu, T. Two-stage spatial temporal deep learning framework for functional brain network modeling. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1576–1580 (2019)
3. Hu, X., et al.: Latent source mining in FMRI via restricted Boltzmann machine. Hum. Brain Mapp. **39**(6), 2368–2380 (2018)

4. Huang, H., et al.: Modeling task fMRI data via deep convolutional autoencoder. IEEE Trans. Med. Imaging **37**(7), 1551–1561 (2017)
5. Zhang, W., et al.: Hierarchical organization of functional brain networks revealed by hybrid spatiotemporal deep learning. Brain Connect. **10**(2), 72–82 (2020)
6. Zhang, W., Zhao, L., Li, Q., Zhao, S., Dong, Q., Jiang, X., et al.: Identify hierarchical structures from task-based fMRI data via hybrid spatiotemporal neural architecture search net. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2019), pp. 745–753 (2019)
7. Wang, H., et al.: Recognizing brain states using deep sparse recurrent neural network. IEEE Trans. Med. Imaging **38**(4), 1058–1068 (2018)
8. Li, Q., Dong, Q., Ge, F., Qiang, N., Zhao, Y., Wang, H., et al.: Simultaneous spatial-temporal decomposition of connectome-scale brain networks by deep sparse recurrent auto-encoders. In: International Conference on Information Processing in Medical Imaging (IPMI 2019), pp. 579–591 (2019)
9. Li, Q., Zhang, W., Lv, J., Wu, X., Liu, T.: Neural architecture search for optimization of spatial-temporal brain network decomposition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2020), pp. 377–386 (2020)
10. Liu, H., Simonyan, K., Yang, Y.: Darts: differentiable architecture search. arXiv preprint arXiv:1806.09055 (2018)
11. Yan, S., Zheng, Y., Ao, W., Zeng, X., Zhang, M.: Does unsupervised architecture representation learning help neural architecture search?. arXiv preprint arXiv:2006.06936 (2020)
12. Li, Q., Wu, X., Liu, T.: Differentiable neural architecture search for optimal spatial/temporal brain function network decomposition. Med. Image Anal. **69**, 101974 (2021)
13. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks?. arXiv preprint arXiv:1810.00826 (2018)
14. Zoph, B., Le, Q.V. Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578 (2016)
15. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. **8**(3–4), 229–256 (1992)
16. White, C., Neiswanger, W., Savani, Y.: Bananas: Bayesian optimization with neural architectures for neural architecture search. arXiv preprint arXiv:1910.11858 (2019)
17. Gardner, M.W., Dorling, S.R.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. **32**(14–15), 2627–2636 (1998)
18. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. **67**(2), 301–320 (2005)
19. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)



Driving Points Prediction for Abdominal Probabilistic Registration

Samuel Joutard(✉), Reuben Dorent, Sebastien Ourselin, Tom Vercauteren, and Marc Modat

King's College London, London, UK
samuel.joutard@kcl.ac.uk

Abstract. Inter-patient abdominal registration has various applications, from pharmakinematic studies to anatomy modeling. Yet, it remains a challenging application due to the morphological heterogeneity and variability of the human abdomen. Among the various registration methods proposed for this task, probabilistic displacement registration models estimate displacement distribution for a subset of points by comparing feature vectors of points from the two images. These probabilistic models are informative and robust while allowing large displacements by design. As the displacement distributions are typically estimated on a subset of points (which we refer to as driving points), due to computational requirements, we propose in this work to learn a driving points predictor. Compared to previously proposed methods, the driving points predictor is optimized in an end-to-end fashion to infer driving points tailored for a specific registration pipeline. We evaluate the impact of our contribution on two different datasets corresponding to different modalities. Specifically, we compared the performances of 6 different probabilistic displacement registration models when using a driving points predictor or one of 2 other standard driving points selection methods. The proposed method improved performances in 11 out of 12 experiments.

1 Introduction

Medical image registration (MIR) is ubiquitous in medical engineering analytical pipelines. While there are many widely used MIR tools [17,26], inter-patient abdominal registration remains particularly challenging. The main challenge of this task is the high geometrical and appearance variabilities of the abdominal region of the human body. These result from the combination of intra-patient variability mainly due to the presence of soft organs highly impacted by the patient's position and current states (breathing, digestion, etc.) and inter-patient abdominal variance. Comparison studies [15,27] suggest two research directions for this task: first, Xu et al. [27] showed that probabilistic displacement methods tend to be less prone to local minima as opposed to other optimization-based MIR algorithms. Probabilistic displacement methods consider a probabilistic distribution

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_30.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 288–297, 2022.
https://doi.org/10.1007/978-3-031-21014-3_30

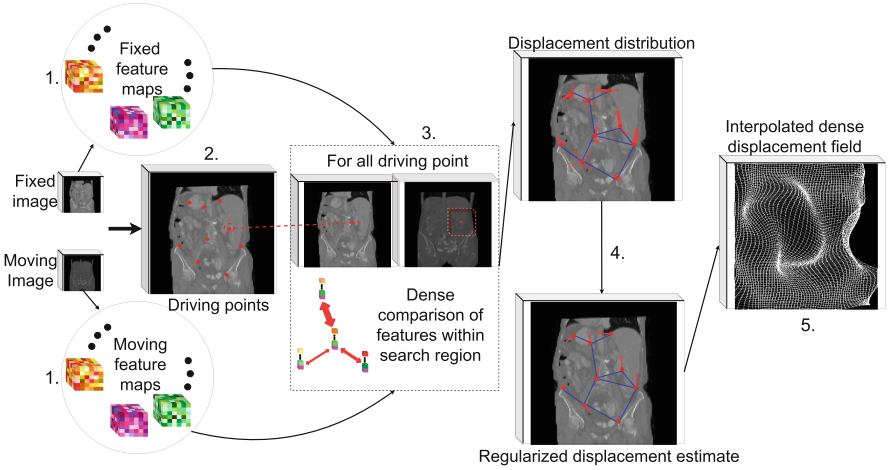


Fig. 1. Generic probabilistic displacement registration pipeline. *1. Dense Feature Extraction*: Dense feature maps are extracted from both the moving and the fixed image. *2. Driving Points Identification*: A subset of points from the fixed image are identified as driving points. *3. Displacement Distribution Estimation*: A displacement distribution is obtained for each driving points by comparing its associated feature vector with and feature vectors from the moving feature maps within a predefined search region. *4. Regularization*: a graph connecting the driving points is used to estimate a regularized displacement field. *5. Interpolation*: the displacement field is interpolated to the original voxel grid.

over deformations derived from local points matching distribution as illustrated in Fig. 1. Secondly, the Learn2Reg (L2R) challenge [15] has demonstrated the superiority of learning-based approaches on this task. Indeed, learning-based regressive models such as LapIRN [18] have shown great success for inter-patient abdominal registration (L2R task 3). The learning-based counter-parts of probabilistic displacement approaches [7,9] achieved great performances on US-MR brain registration, lung registration, and intra-patient abdominal image fusion but did not reach the same level of performance on inter-patient abdominal registration. Yet, these approaches allow for large deformations by design as they evaluate displacement probabilities within a (large) specified search region. Also, as probabilistic approaches, they should be more robust to the noise and ambiguity inherent to inter-patient abdominal registration.

Probabilistic displacement approaches typically have a high memory and computational cost due to the evaluation of displacement probabilities within a search region. Hence, these methods commonly operate on a subset of driving points (step 2. in Fig. 1) which are a lower resolution grid [4,9,13,20], uniformly sampled random points [8] or key-points [7,10,21,23,24,28]. These approaches disentangle the selection of the driving points from the rest of the pipeline. Yet, this step impacts the matching process (driving points should be matched unambiguously) as well as the deformation complexity (the driving points should be placed to capture the relevant variations of the deformation field). In this

work, we proposed to learn a driving points predictor in an end-to-end fashion. As such, the driving points prediction is integrated and optimized alongside other learnable components and is specifically tailored to a particular probabilistic displacements registration pipeline. For natural image matching and perception, landmarks are usually obtained as local maxima of a response map [25,30]. This response map can be handcrafted or optimized as a neural network. We proposed instead to obtain the driving points by predicting a deformation field to be applied to a low-resolution regular grid. The vertices of the deformed regular grid are then considered as driving points. This format ensures a minimum coverage of the image and a fixed amount of driving points. It is also a memory-efficient prediction format that still allows predicting locations with sub-voxel precision. We conducted experiments on two datasets, a CT dataset from the L2R 2020 challenge [27] and the CHAOS MR dataset [16]. We demonstrate improvement in most probabilistic displacement models when using a driving points predictor against two other standard driving points selection methods.

2 Method

In this section, we first describe the five steps of probabilistic displacement registration pipelines (see Fig. 1) and their associated methods from the literature before introducing in detail the proposed driving points prediction method which can be used for the second step to improve the overall pipeline performances.

2.1 Pipeline Overview

1. Dense Feature Extraction. The objective of the dense feature extractor is to associate a vector characterizing the corresponding anatomical point to each voxel. Various handcrafted feature extraction methods have been suggested. The most straightforward option is to rely on image intensities [1–4,12] and/or intensity gradients [2,13,23]. Wavelet decomposition [28] and 3D Gabor attributes [19] have been used to enrich these feature vectors. Other methods proposed to describe voxels using local self-similarity measurements such as the self-similarity context [14] or MIND descriptors [7,8,10]. Finally, recent models adopted a data-driven approach by learning deep feature extractors [6,9,11].

2. Driving Points Identification. Instead of matching all points, probabilistic displacement models typically focus on a subset of points called driving points. This helps to reduce the computational cost. If driving points are well selected, it also mitigates the impact of points located in ambiguous regions and, thus, difficult to match accurately. The most straightforward approach is to rely on a low resolution regular grid [4,5,9,13,20] or to use random points drawn from a uniform distribution [8]. More sophisticated approaches rely on key-points extractors such as the Foerstner key-point extractor [7,10,23], 3D-SIFT points [21] or handcrafted key-points extractors [24,28,29]. Key-points extractor based on the image input only (no additional segmentation provided) focus on salient regions

in the fixed image by considering local maxima of a saliency function. Yet, for abdominal image MIR, this does not necessarily mean that the point can be unambiguously matched or is of interest for the deformation. Another drawback of such methods is the lack of guarantee to cover sufficiently the regions of interest in the image. In contrast, the proposed driving points predictor presented in the following section covers by design a large region of the abdomen and optimizes a typical registration objective term.

3. Displacement Distribution Estimation. The next step estimates a displacement distribution for each driving point. Specifically, a matching probability is derived from the Euclidean distance between the feature vectors describing each driving point and the feature vectors of matching candidates. In general, matching candidates are assumed to be in a pre-defined window centered around the driving point [7,9]. Alternatively, random displacements sampled from uniform distributions can be used [8,12].

4. Regularization. The regularization step promotes smooth deformations, which are assumed to be more realistic. In order to promote spatially coherent deformations, several options have been proposed such as fitting a transformation model having low degree of freedom to the matching distribution [1]. Regarding probabilistic regularization models, deformations are typically assumed to be drawn from a Markov Random Field (MRF) distribution. The individual displacement distributions estimated at the previous step define the unary term of the MRF distribution, while the pairwise term promote local displacement coherence. The maximum a posteriori (MAP) estimate of this distribution can be computed using the Fast-PD algorithm [4,5,19,20]. As this algorithm introduces a high memory and computational cost, the resolution of the MRF distribution can be approximated by replacing the neighboring graph with a minimum spanning tree to compute the MAP estimator in a closed form [10,12–14,23]. Alternatively, Mean Field (MF) approximation can be employed to approximate the mean estimator or the MAP estimator [8,9,11]. Finally, regularization can be performed using a neural network [6,7].

5. Interpolation. The last step consists of interpolating the deformation back to the original resolution. Standard approaches rely on non-overlapping [2,24,28] and overlapping [3] Gaussian kernel interpolation or linear interpolation [9].

2.2 Driving Points Prediction

Driving point sets should correspond to points which can be reliably matched within their associated search region in the moving image. For this reason, the driving points selection should be adapted to the feature extraction step. Ideally, the driving point set should be specific to a fixed/moving image pair rather than only depending on the fixed image as the ambiguity of the matching process depends on both inputs. Finally, the driving points selection impacts the subsequent steps of the pipeline so it should be adapted to the pipeline as a

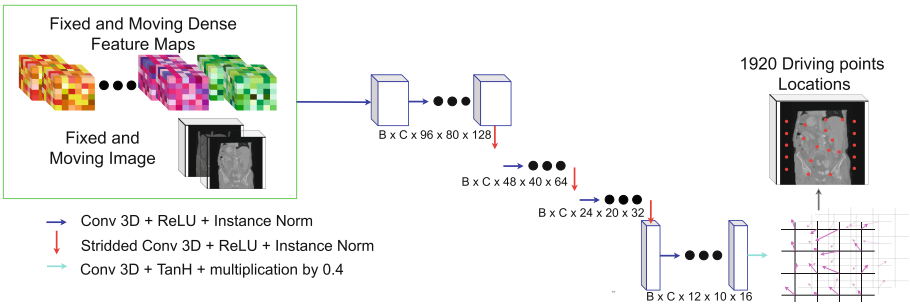


Fig. 2. Driving points predictor architecture

whole, just like the pipeline should be adapted to it. For this reason, we propose to learn the driving points selection process in an end-to-end fashion to select points that are adapted to the rest of the pipeline to optimize the registration objective.

The proposed architecture for the driving points predictor is illustrated in Fig. 2. The driving points predictor takes as input the concatenation of both the fixed and the moving image as well as their associated dense feature maps. Hence, the driving points predictor can exploit the specific feature maps landscape to predict relevant points to match. The input tensor goes through an encoder network with several downsampling layers to obtain a lower resolution embedding of the input tensor. From there, directly predicting the driving points' locations (by flattening the low-resolution embedding and passing it through a set of fully connected layers for instance) is particularly challenging as it is an unstructured prediction process very sensitive to the random initialization. Instead, we propose the use of a structured output by deforming a low-resolution regular grid. Indeed, as illustrated in Fig. 2, the final layer of the driving points predictor is also convolutional and predicts a displacement field to be applied to the low-resolution regular grid. The vertices of the deformed low-resolution grid are considered as driving points. To ensure a minimum coverage of the image, we cap the maximum displacement range to 20% of the image dimension in all directions. The number of predicted driving points typically corresponds to the number G of vertices in the low-resolution regular grid. In theory, multiple vertices could be mapped to the same points, leading to fewer driving points but we never observed that phenomenon in practice. In case more driving points are required to improve performances, one could either use a higher resolution grid (i.e. use fewer downsampling layers) or make the last layer predict D different displacement fields to obtain $D \times G$ driving points.

Once the driving points' locations have been predicted, the fixed dense feature map is sampled at the predicted locations using linear interpolation to obtain the descriptors of the driving points. This step is differentiable and allows to learn the driving points predictor parameters through back-propagation. The parameters of the driving points predictor are learned simultaneously with other learnable components of the pipeline (e.g. the feature extractor or the regularization network) by minimizing a standard registration loss, which in our experiments is

a combination of local normalized cross-correlation and a Hessian penalty term (mathematical formulation in supplementary material).

3 Experiments

3.1 Datasets

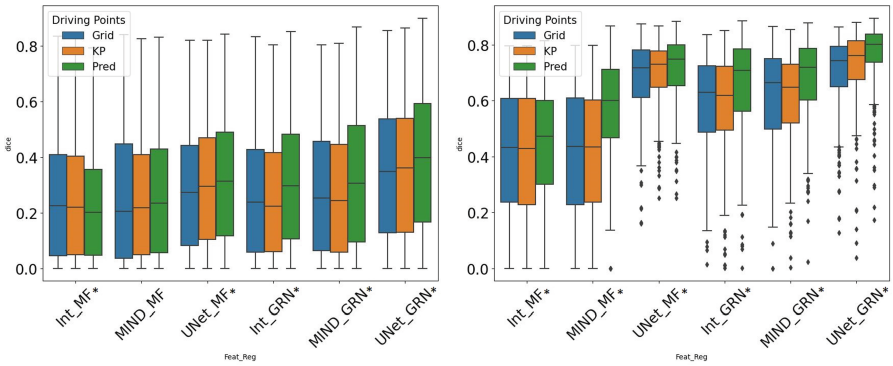
We conducted experiments on two different datasets. First, we used the combined training and validation sets of the L2R 2020 challenge task 3 [15]. This dataset contains 30 Abdominal CT volumes. These volumes also come with the segmentation of 13 structures: the spleen, the left/right kidney, the gallbladder, the esophagus, the liver, the stomach, the aorta, the inferior vena cava, the portal and splenic veins, the pancreas, the left/right adrenal gland. Then, we considered the Magnetic Resonance (MR) volumes of the CHAOS dataset [16]. This dataset contains 40 MR volumes from different patients with 4 structures annotated: the spleen, the left/right kidney, and the liver. For each dataset, we saved 10 volumes for testing (i.e. 90 pairs of patients to be registered) and used the rest of the volumes for training/validation. We refer from now on to the L2R challenge dataset as "CT dataset" and to the Chaos dataset as "MR dataset".

3.2 Experimental Setting

As detailed in Sect. 2.1, various solutions have been proposed for each step. In this work, we introduced a new data-driven and feature-aware driving points selection process. To assess the effectiveness of the proposed method, we learned driving points predictors as part of different pipelines. Specifically, we considered three different dense feature extractors: the image intensity as a baseline, MIND descriptors [7,8,10] as a state-of-the-art option for handcrafted features for MIR or feature vectors extracted using a standard UNet architecture [22]. We then compared our driving points predictor with two traditional driving points selection approaches based on a low-resolution grid and Foerstner key-points. Regarding the displacement distribution estimation, we used a predefined search region associated to the set of admissible displacements $\{-25, -20, ...20, 25\}^3$ (expressed in voxels). Two forms of regularization were tested, either based on mean-field approximation of an MRF formulation as in PDDNet [9] or a neural network graph-based approach as in GraphRegNet (GRN) [7]. Linear interpolation was used for the interpolation step.

To assess the quality of the registration, we report Dice scores for each annotated structure. We also report regularity metrics in the supplementary material. All learnable components were trained in an end-to-end fashion by minimizing a standard MIR loss combining the local normalized cross-correlation loss as similarity metric and a regularization term penalizing the norm of the Hessian of the displacement field. Not using label information during training allows a fair comparison with methods that have very limited learned parameters. All hyperparameters were optimized using grid-search and are available in our code[1].

[1] https://github.com/SamuelJoutard/DrivingPointsPredictionMIR.



(a) CT dataset (b) MR dataset

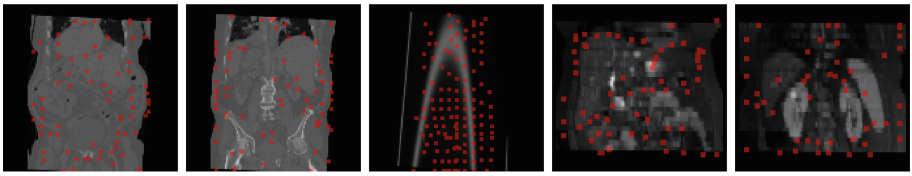
Fig. 3. Mean Dice scores for the considered probabilistic displacement registration pipelines. Grid, KP and Pred respectively correspond to using a low resolution grid, Foerstner Key-points and the proposed driving points predictor. The x-axis indicates the feature/regularization configuration. (*) indicates configurations where the ranking is statistically significant (Wilcoxon rank test, $p < 0.01$). (a) CT dataset. (b) MR dataset.

3.3 Results

Averaged Dice scores are reported in Fig. 3 and more detailed results are available in Tables S1 and S2 of supplementary material. Our method significantly improved the registration compared to using Foerstner key-points or a low resolution grid for 10 of the 12 Features/Regularization/Dataset configurations tested. This highlights the effectiveness of our approach. The improvement observed for the MIND/Mean-Field/CT dataset is not statistically significant by the Wilcoxon test and we observed a performance downgrade for the intensity/mean-field/CT dataset due to some overfitting we did not manage to prevent. We also note that our approach leads to large improvements on the MR dataset when using the intensity or MIND features which shows that optimizing the driving points for a specific dense feature extraction method is particularly effective when the feature extraction step is sub-optimal. The small structures segmented in the CT dataset such as the gallbladder are particularly challenging to align which explains the lower mean Dice scores observed on this dataset.

We illustrate some of the predicted driving points in Fig. 4. First, we observe by comparing Fig. 4c to the other slices shown that the regular grid structure is only preserved far from the regions of interest. The driving points predictor selected specific locations within the fixed patient's abdomen. We can also observe that regions that are hard to match like the intestinal region (see Fig. 4a) or regions with low contrast like the interior of the liver (Fig. 4b–d) are not densely populated with driving points. This suggests that the registration pipeline tends to focus on informative points within the patient when using our method.

Finally, to evaluate the respective contributions of the fixed and moving images in the driving points selection, we compared predicted driving point sets.



(a) (b) (c) (d) (e)

Fig. 4. Predicted driving points represented in red overlapping the fixed image. (a, b, c) show coronal slices of the same patient from the CT dataset. (c) shows driving points predicted far from the regions of interest. (d, e) show coronal slices of the same patient from the MR dataset. (Color figure online)

Specifically, we computed the Wassertein-2 distance (W2) between predicted driving point sets for different inputs. We compared the mean W2 distance between predicted driving point sets used to register two pairs of images sharing the same fixed image against the mean W2 distance between all predicted driving point sets. The former is on average 7 times smaller than the latter. This suggests that the predicted driving point sets are specific to the fixed image. More detailed comparisons are available in Fig. S6 of supplementary materials. We also compared predicted driving points for the same input pair of images to be registered but using different driving points predictors trained for different pipelines. We obtained mean W2 distance that ranges from 13 to 41 for the CT dataset and from 11 to 29 for the MR dataset. This confirms that the predicted driving point sets are pipeline specific.

4 Discussion

In this work, we proposed an encoder architecture to predict driving point set for probabilistic displacements MIR models applied to abdominal images. The proposed architecture is stable to train, ensures a certain coverage of the fixed image, and has a memory-saving prediction format. We showed that predicting driving points in this way yielded performance improvements for almost all configurations tested which involved two different datasets. Finally, we showed that the proposed architecture was able to predict driving point sets tailored to the fixed image data and to the specific steps of the registration pipeline.

The presented methods did not integrate all the potential refinements that have been shown to improve performances such as multiple forward passes with intermediate wrapping or instance optimization. Similarly, in this work, the segmentation masks were used for evaluation purposes only. Given that previous studies have shown that exploiting these segmentation masks during training leads to performance improvement, we plan to integrate a segmentation matching loss alongside other model improvements in future work to further improve our results.

References

1. Castillo, E., Castillo, R., Fuentes, D., Guerrero, T.: Computing global minimizers to a constrained B-spline image registration problem from optimal l 1 perturbations to block match data. Med. Phys. **41**(4), 041904 (2014)
2. Foroughi, P., Abolmaesumi, P.: Elastic registration of 3D ultrasound images. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 83–90. Springer, Heidelberg (2005). https://doi.org/10.1007/11566465_11
3. Garcia, V., Commowick, O., Malandain, G.: A robust and efficient block matching framework for non linear registration of thoracic CT images. In: Grand Challenges in Medical Image Analysis (MICCAI Workshop) (2010)
4. Glocker, B., Komodakis, N., Tziritas, G., Navab, N., Paragios, N.: Dense image registration through MRFs and efficient linear programming. Med. Image Anal. **12**, 731–741 (2008)
5. Glocker, B., Sotiras, A., Komodakis, N., Paragios, N.: Deformable medical image registration: setting the state of the art with discrete methods. Annu. Rev. Biomed. Eng. **13**, 219–244 (2011)
6. Ha, I.Y., Heinrich, M.P.: Comparing deep learning strategies and attention mechanisms of discrete registration for multimodal image-guided interventions. In: Zhou, L., et al. (eds.) LABELS/HAL-MICCAI/CuRIOUS -2019. LNCS, vol. 11851, pp. 145–151. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33642-4_16
7. Hansen, L., Heinrich, M.P.: GraphRegNet: deep graph regularisation networks on sparse keypoints for dense registration of 3D lung CTs. IEEE Trans. Med. Imaging **40**, 2246–2257 (2021)
8. Hansen, L., Heinrich, M.P.: Revisiting iterative highly efficient optimisation schemes in medical image registration. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 203–212. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_20
9. Heinrich, M.P.: Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 50–58. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_6
10. Heinrich, M.P., Handels, H., Simpson, I.J.A.: Estimating large lung motion in COPD patients by symmetric regularised correspondence fields. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9350, pp. 338–345. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24571-3_41
11. Heinrich, M.P., Hansen, L.: Highly accurate and memory efficient unsupervised learning-based discrete CT registration using 2.5D displacement search. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 190–200. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_19
12. Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A.: MRF-based deformable registration and ventilation estimation of lung CT. IEEE Trans. Med. Imaging **32**, 1239–1248 (2013)
13. Heinrich, M.P., Simpson, I.J.A., Jenkinson, M., Brady, M., Schnabel, J.A.: Uncertainty estimates for improved accuracy of registration-based segmentation propagation using discrete optimisation (2013)
14. Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, S.M., Schnabel, J.A.: Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 187–194. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40811-3_24

15. Hering, A., et al.: Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning (2021)
16. Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S.: CHAOS - combined (CT-MR) healthy abdominal organ segmentation challenge data (2019)
17. Modat, M., et al.: Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed. **98**, 278–284 (2009)
18. Mok, T.C.W., Chung, A.C.S.: Large deformation diffeomorphic image registration with Laplacian pyramid networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 211–221. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_21
19. Ou, Y., Sotiras, A., Paragios, N., Davatzikos, C.: DRAMMS: deformable registration via attribute matching and mutual-saliency weighting. Med. Image Anal. **15**, 622–639 (2011)
20. Parisot, S., Wells, W., Chemouny, S., Duffau, H., Paragios, N.: Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs. Med. Image Anal. **18**, 647–659 (2014)
21. Rister, B., Horowitz, M., Rubin, D.: Volumetric image registration from invariant keypoints. IEEE Trans. Image Process. **27**, 4900–4910 (2017)
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
23. Ruhaak, J., et al.: Estimation of large motion in lung CT by integrating regularized keypoint correspondences into dense deformable registration. IEEE Trans. Med. Imaging **36**, 1746–1757 (2017)
24. Shen, D., Davatzikos, C.: HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imaging **21**, 1421–1439 (2002)
25. Thewlis, J., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks by factorized spatial embeddings. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), October 2017
26. Tustison, N.J., et al.: Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. Neuroimage **99**, 166–179 (2014)
27. Xu, Z., et al.: Evaluation of six registration methods for the human abdomen on clinically acquired CT. IEEE Trans. Biomed. Eng. **63**, 1563–1572 (2016)
28. Xue, Z., Shen, D., Davatzikos, C.: Determining correspondence in 3-D MR brain images using attribute vectors as morphological signatures of voxels. IEEE Trans. Med. Imaging **23**, 1276–1291 (2004)
29. Yap, P.T., Wu, G., Zhu, H., Lin, W., Shen, D.: TIMER: tensor image morphing for elastic registration. Neuroimage **47**, 549–563 (2009)
30. Zhang, Y., Guo, Y., Jin, Y., Luo, Y., He, Z., Lee, H.: Unsupervised discovery of object landmarks as structural representations. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)



CircleSnake: Instance Segmentation with Circle Representation

Ethan H. Nguyen[1(✉)], Haichun Yang[2], Zuhayr Asad[1], Ruining Deng[1], Agnes B. Fogo[2], and Yuankai Huo[1]

[1] Vanderbilt University, Nashville 37235, USA
{ethan.h.nguyen,yuankai.huo}@vanderbilt.edu
[2] Vanderbilt University Medical Center, Nashville 37215, USA

Abstract. Circle representation has recently been introduced as a "medical imaging optimized" representation for more effective instance object detection on ball-shaped medical objects. With its superior performance on instance detection, it is appealing to extend the circle representation to instance medical object segmentation. In this work, we propose CircleSnake, a simple end-to-end circle contour deformation-based segmentation method for ball-shaped medical objects. Compared to the prevalent DeepSnake method, our contribution is threefold: (1) We replace the complicated *bounding box to octagon contour* transformation with a computation-free and consistent *bounding circle to circle contour* adaption for segmenting ball-shaped medical objects; (2) Circle representation has fewer degrees of freedom (DoF = 2) as compared with the octagon representation (DoF = 8), thus yielding a more robust segmentation performance and better rotation consistency; (3) To the best of our knowledge, the proposed CircleSnake method is the first end-to-end circle representation deep segmentation pipeline method with consistent circle detection, circle contour proposal, and circular convolution. The key innovation is to integrate the circular graph convolution with circle detection into an end-to-end instance segmentation framework, enabled by the proposed simple and consistent circle contour representation. Glomeruli are used to evaluate the performance of the benchmarks. From the results, CircleSnake increases the average precision of glomerular detection from 0.559 to 0.614. The Dice score increased from 0.804 to 0.849. The code has been released: https://github.com/hrlblab/CircleSnake.

Keywords: Instance segmentation · Graph convolution · Pathology · Snake

1 Introduction

Bounding box and polygon representations from the computer vision community have been widely utilized for medical object detection [6,9,13,18]. However, such representations are not necessarily optimized (Fig. 1) for medical objects. Unlike natural images, certain biomedical images (e.g., microscopy imaging can be obtained and displayed at any angle of rotation of the same tissue [19]. Thus,

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 298–306, 2022.
https://doi.org/10.1007/978-3-031-21014-3_31

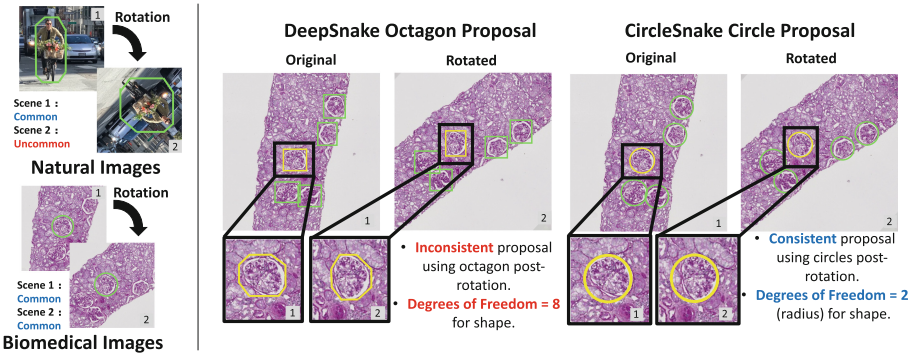


Fig. 1. Comparison of traditional representation and the circle representation. The left panel shows that samples of glomeruli can be scanned at any angle of rotation. The right panel highlights how the octagon proposal is not optimized for the ball-shaped glomerulus. Using the proposed CircleSnake yields a more consistent representation while using fewer degrees of freedom (DoF). Moreover, the proposed circle contour proposal is fundamentally easier and more consistent as compared with the computer vision-oriented octagon contour proposal (Fig. 2).

traditional object representation might yield inferior performance for representing such ball-shaped objects [15].

Recently, circle representation has been introduced as a "medical imaging optimized" representation for instance object detection on ball-shaped medical objects, such as glomeruli, nuclei, and tumors [14,15,19]. Given its superior performance, it is appealing to extend the circle representation to the instance object segmentation. In renal pathology, the ability to precisely segment glomeruli is critical to investigating several kidney diseases [2,7]. Most prior arts [1,3,4,8] are pixel-based where they perform instance segmentation within a regional proposal on the pixel level. A popular example, Mask R-CNN [5], first detects objects and then segments instances within the proposed boxes using a mask predictor. In contrast, DeepSnake [16] addresses localization errors from the detector by deforming the detected bounding octagon to object boundaries. However, DeepSnake's bounding octagon is not optimal for the ball-shaped glomerulus (Fig. 1) and furthermore, is computationally complicated (Fig. 2).

In this paper, we propose a contour-based approach that utilizes the circle representation called CircleSnake for the robust segmentation of glomeruli. The "bounding circle" is introduced as the detection and initial contour representation for the ball-shaped structure. Once the center location of the lesion is obtained, only degrees of freedom (DoF) = 2 is required to form the bounding circle, while DoF = 8 is required for the bounding octagon. Briefly, the contributions of this study are in three areas:

- **Optimized Biomedical Object Segmentation**: To the best of our knowledge, the proposed CircleSnake is the first contour deformation-based end-to-end segmentation approach for detecting ball-shaped medical objects.

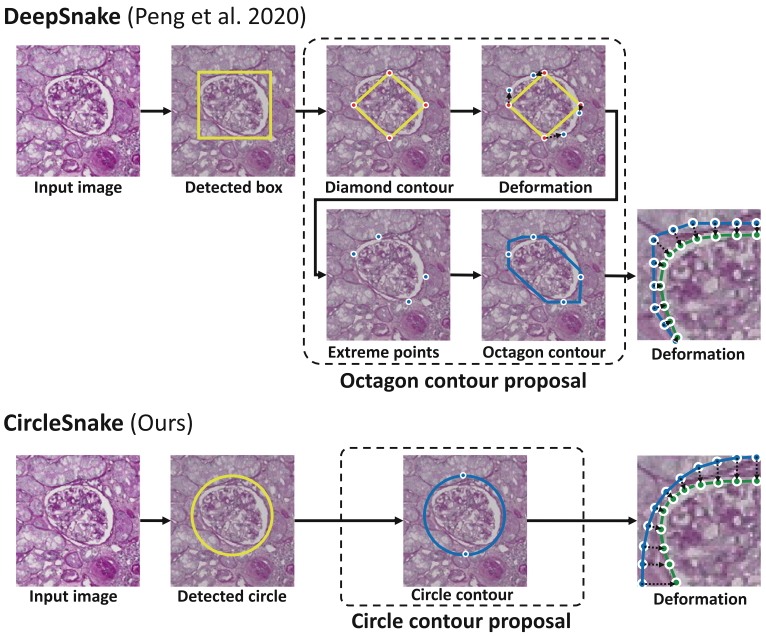


Fig. 2. Circle contour proposal. This figure presents the differences between the "bounding box to octagon contour" representation and the proposed "bounding circle to circle contour" representation. Our circle contour proposal avoids the complicated extreme points and deformation-based contour proposal by directly introducing a simple circle proposal, which bridges the circle detection and deformation-based segmentation in an computational-free and end-to-end manner.

- **Circle Representation**: We propose a consistent circle representation pipeline for segmenting ball-shaped biomedical objects with integrated (1) circle detection (2) circle contour proposal, and (3) circular convolution, with a smaller DoF of fitting and superior segmentation performance.
- **Rotation Consistency**: The proposed circle representation yields fewer DoF of fitting, better segmentation performance, and superior rotation consistency.

2 Methods

Figure 3 provides an overview of the proposed pipeline for instance segmentation. Inspired by [16], the object segmentation is implemented by deforming an initial rough contour to match object boundary. Different from DeepSnake which uses bounding box representation and polygon contour, we propose to use circle detection [15] and circle contour proposal (Fig. 1) in our CircleSnake method. Then the proposed circle initial contour is deformed via circular convolution by predicting the per-vertex offsets towards the real object boundary (Fig. 2). Such deformation is performed iteratively to obtain the final object shape as instance segmentation.

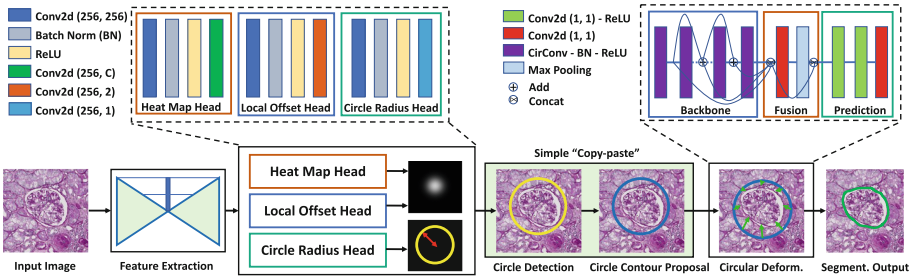


Fig. 3. The network structure of proposed CircleSnake. The proposed CircleSnake After detection by CircleNet, the bounding circle forms the initial contour. This contour is deformed into the final contour using a graph convolutional network (GCN).

2.1 Circle Object Detection

The circle object detection follows the design of the recent CircleNet approach [15,19] for its high performance and simplicity (Fig. 3). Following the definitions from Zhou *et al.* [21], let $I \in R^{W \times H \times 3}$ be an input image with height H and width W. The CPL network produces the center localization of each object within a heatmap $\hat{Y} \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times C}$ where C is the number of candidate class and R is a downsampling factor. Within the heatmap, $\hat{Y}_{xyc} = 1$ is the center of the lesion and $\hat{Y}_{xyc} = 0$ is the background. Following convention, [10,21], the target center point is splat on a heatmap as a 2D Gaussian kernel:

$$Y_{xyc} = \exp\left(-\frac{(x-\tilde{p}_x)^2 + (y-\tilde{p}_y)^2}{2\sigma_p^2}\right) \tag{1}$$

where the $\tilde{p}_x$ and $\tilde{p}_y$ are the downsampled ground truth center points and σ_p is the kernel standard deviation. The training loss L_k is penalty-reduced logistic regression with focal loss [11]:

$$L_k = \frac{-1}{N} \sum_{xyc} \begin{cases} (1-\hat{Y}_{xyc})^{\alpha} \log(\hat{Y}_{xyc}) & \text{if } Y_{xyc} = 1 \\ (1-Y_{xyc})^{\beta} (\hat{Y}_{xyc})^{\alpha} & \\ \log(1-\hat{Y}_{xyc}) & \text{otherwise} \end{cases} \tag{2}$$

where the hyper-parameters α and β are kept the same as [11]. To further refine the prediction location, the ℓ_1-norm offset prediction loss L_{off} is used.

To obtain the center point, the top n peaks whose values are greater or equal to its 8-connected neighbors are proposed. These n detected center points are defined as $\hat{\mathcal{P}} = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{n}$. The center point of each object contains an integer coordinate (x_i, y_i) from $\hat{Y}_{x_i y_i c}$ and L_k. The offset $(\delta\hat{x}_i, \delta\hat{y}_i)$ is obtained from L_{off}. The bounding circle is combines the center point $\hat{p}$ and radius $\hat{r}$ as:

$$\hat{p} = (\hat{x}_i + \delta\hat{x}_i, \quad \hat{y}_i + \delta\hat{y}_i). \quad \hat{r} = \hat{R}_{\hat{x}_i, \hat{y}_i}. \tag{3}$$

where $\hat{R} \in \mathcal{R}^{\frac{W}{R} \times \frac{H}{R} \times 1}$ contains the radius prediction for each pixel, optimized by

$$L_{radius} = \frac{1}{N} \sum_{k=1}^{N} \left| \hat{R}_{p_k} - r_k \right|. \tag{4}$$

where r_k is the ground truth radius for each object k. Finally, the overall objective is

$$L_{det} = L_k + \lambda_{radius} L_{radius} + \lambda_{off} L_{off}. \tag{5}$$

Following [21], we set $\lambda_{radius} = 0.1$ and $\lambda_{off} = 1$.

2.2 Circle Contour Proposal

Using the CircleNet based object detection [19], bounding circle based object detection is obtained for every target object. Then, the initial circle contour proposal is directly achieved from the circle representation (Fig. 2). Our new circle contour proposal skips the the original complicated deformation and extreme point-based octagon contour proposal; ultimately, this makes the contour proposal both simple and consistent. The circle proposal is determined by the center point and radius. N initial points $\{\mathbf{x}_i^{circle} | i = 1, 2, ..., N\}$ are uniformly sampled from the circle contour starting at the top-most point x_1^{circle} . Similarly, the ground truth contour is formed by sampling N vertices clockwise along the object boundary. The N is set to 128 based on [16].

2.3 Circular Contour Deformation

Given a contour with N vertices $\{\mathbf{x}_i^{circle} | i = 1, ..., N\}$, we first construct feature vectors for each vertex. The input feature f_i^{circle} for a vertex $\mathbf{x}_i^{circle}$ is a concatenation of learning-based features and the vertex coordinate: $[F(\mathbf{x}_i^{circle}); \mathbf{x}_i^{circle}]$, where F denotes the feature maps. The input features defined are treated as a 1-D discrete signal $f : \mathbb{Z} \rightarrow \mathbb{R}^D$ on a circle contour. We utilize the circular convolution for the feature learning, as illustrated [16]. The f is defined as a periodic signal as:

$$(f_N^{circle} * k)_i = \sum_{j=-r}^{r} (f_N^{circle})_{i+j} k_j, \tag{6}$$

where $k : [-r, r] \rightarrow \mathbb{R}^D$ is a learnable kernel function, while the operator $*$ is the standard convolution. Following [16], the kernel size of the circular convolution is fixed to be nine.

In CircleSnake, the above convolution is implemented via the graph convolutional network (GCN) inspired by [16]. The GCN consists of three parts: backbone, fusion, and prediction. The backbone contains eight "CirConv-Bn-ReLU" layers with residual skip connections and "CirConv" for circular convolution. The fusion block combines the information across contour points at different scales. Specifically, the features are concatenated from all layers in the backbone and forwarded through a 1×1 convolutional layer and max pooling layer.

Finally, the prediction head contains three 1×1 convolution layers and outputs vertex-wise offsets. The loss function for the iterative contour deformation is defined as

$$L_{iter} = \frac{1}{N} \sum_{i=1}^{N} l_1(\tilde{x}_i^{circle} - x_i^{gt}). \tag{7}$$

where x_i^{gt} is the ground truth boundary point and $\tilde{x}_i^{circle}$ is the deformed contour point. Following [16], we regress the N offsets in 3 iterations.

3 Experimental Design

For baseline methods, DeepSnake [16] was utilized for their superior performance in instance segmentation. Deep layer aggregation (DLA) network [20] were employed as backbone networks. The implementations the backbone networks and segmentation networks followed the authors' official PyTorch implementations. All models were initialized with the COCO pretrained model [12]. The same workstation with an Nvidia 3090 was used to perform all experiments in this study. For both the glomeruli and nuclei experiments, the hyperparameters were set to maximum epoch = 50, learning rate = $1e-4$, batch size = 16, and optimizer = Adam.

Conventional segmentation metrics were used. For detection, we utilized, average precision (AP), AP_{50} (IOU threshold at 0.5), AP_{75} (IOU threshold at 0.75), AP_S (small scale with area $< 32^2$), AP_M (medium scale with area $> 32^2$). For segmentation, Dice score was used. as [15].

4 Results

4.1 Glomeruli Results

To form a glomeruli dataset, whole slide images of renal biopsies were captured and annotated. The kidney tissue was routinely processed, paraffin embedded, and $3\mu m$ thickness sections cut and stained with hematoxylin and eosin (HE), periodic acid-Schiff (PAS) or Jones. The samples were deidentified and these studies were approved by the Institutional Review Board (IRB). The training data contained 704 glomeruli from 42 biopsy samples, the validation data contained 98 glomeruli from 7 biopsy samples, and the testing data contained 147 glomeruli from 7 biopsy samples. Considering the size of a glomerulus [17] and its ratio within an image patch, the original high-resolution whole slide images (0.25 μm per pixel) are downsampled to a lower resolution (4 μm per pixel). Then, 512×512 pixels image patches with at least one glomerulus were randomly sampled. These data ultimately comprised a dataset with 7,040 training, 980 validation, and 1,470 testing images.

Referring to Table 1, CircleSnake produces superior detection performane in all average precision metrics. As seen in Table 2 and Fig. 4, CircleSnake outperforms all baseline methods in Dice score.

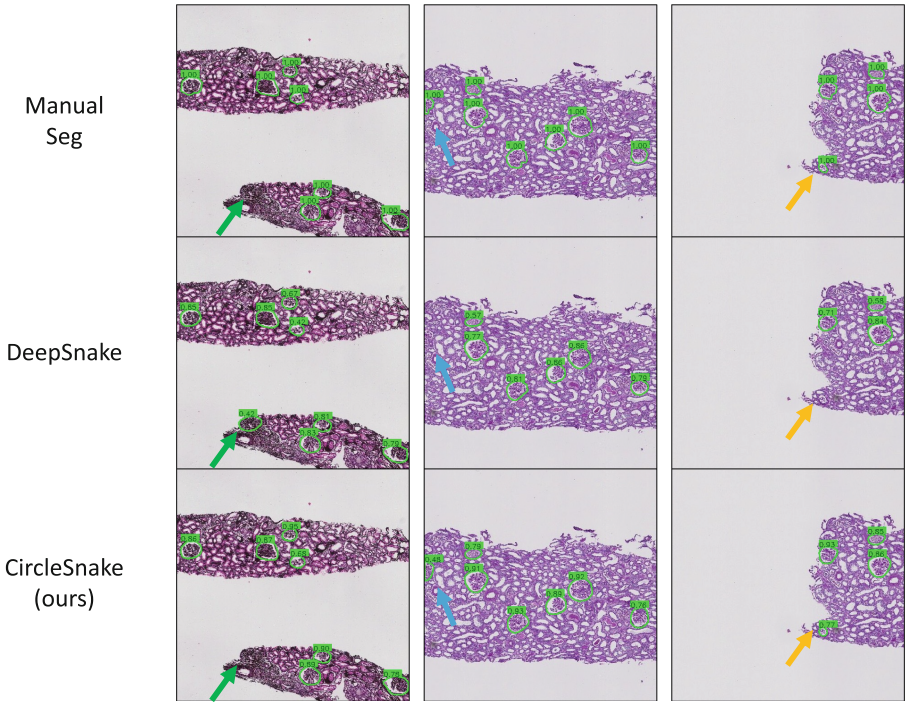


Fig. 4. Qualitative Results. This figure shows the qualitative segmentation results of different methods, compared with manual annotations.

Table 1. Detection and segmentation performance on glomeruli.

Methods	Backbone	AP	$AP_{(50)}$	$AP_{(75)}$	$AP_{(S)}$	$AP_{(M)}$
CenterNet [21]	DLA	0.547	0.872	0.643	0.432	0.640
CircleNet [15]	DLA	0.570	0.857	0.691	0.466	0.647
DeepSnake [16] (Detection)	DLA	0.546	0.880	0.646	0.436	0.634
CircleSnake (Detection) (Ours)	DLA	0.590	0.890	0.701	**0.545**	0.635
DeepSnake [16] (Segmentation)	DLA	0.559	0.874	0.682	0.421	0.660
CircleSnake (Segmentation) (Ours)	DLA	**0.614**	**0.893**	**0.737**	0.544	**0.673**

Table 2. Segmentation performance.

Methods	Backbone	Glomeruli (Dice)
DeepSnake [16]	DLA	0.804
CircleSnake (Ours)	DLA	**0.849**

5 Conclusion

In this paper, we introduce CircleSnake, an circle representation based end-to-end deep learning method for instance medical object segmentation. The proposed CircleSnake seamlessly integrated (1) circle detection (2) circle contour proposal, and (3) circular convolution for segmenting ball-shaped glomeruli, with the superior detectuib instance segmentation performance. The results show that the circle representation does not sacrifice its effectiveness with fewer DoF compared with traditional bounding box and octagon representations. While the circle representation has been implemented with the CircleSnake method, it can be extended to any other instance segmentation method.

Acknowledgements. This work was supported by NIH NIDDK DK56942(ABF).

References

1. Bueno, G., Fernandez-Carrobles, M.M., Gonzalez-Lopez, L., Deniz, O.: Glomerulosclerosis identification in whole slide images using semantic segmentation. Comput. Methods Programs Biomed. **184**, 105273 (2020)
2. D'Agati, V.D., Mengel, M.: The rise of renal pathology in nephrology: structure illuminates function. Am. J. Kidney Dis. **61**(6), 1016–1025 (2013)
3. Gadermayr, M., Dombrowski, A.K., Klinkhammer, B.M., Boor, P., Merhof, D.: CNN cascades for segmenting whole slide images of the kidney. arXiv preprint arXiv:1708.00251 (2017)
4. Ginley, B., Lutnick, B., Jen, K.Y., Fogo, A.B., Jain, S., Rosenberg, A., Walavalkar, V., Wilding, G., Tomaszewski, J.E., Yacoub, R., et al.: Computational segmentation and classification of diabetic glomerulosclerosis. J. Am. Soc. Nephrol. **30**(10), 1953–1967 (2019)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
6. Heckenauer, R., et al.: Real-time detection of glomeruli in renal pathology. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 350–355. IEEE (2020)
7. Huo, Y., Deng, R., Liu, Q., Fogo, A.B., Yang, H.: Ai applications in renal pathology. Kidney International (2021)
8. Kannan, S., Morgan, L.A., Liang, B., Cheung, M.G., Lin, C.Q., Mun, D., Nader, R.G., Belghasem, M.E., Henderson, J.M., Francis, J.M., et al.: Segmentation of glomeruli within trichrome images using deep learning. Kidney international reports **4**(7), 955–962 (2019)
9. Kawazoe, Y.: Faster R-CNN-based glomerular detection in multistained human whole slide images. J. Imaging **4**(7), 91 (2018)
10. Law, Hei, Deng, Jia: CornerNet: Detecting Objects as Paired Keypoints. Int. J. Comput. Vis. **128**(3), 642–656 (2019). https://doi.org/10.1007/s11263-019-01204-1
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988 (2017)
12. Lin, T.-Y., et al.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48

13. Lo, Y.-C.: Glomerulus Detection on Light Microscopic Images of Renal Pathology with the Faster R-CNN. In: Cheng, Long, Leung, Andrew Chi Sing., Ozawa, Seiichi (eds.) ICONIP 2018. LNCS, vol. 11307, pp. 369–377. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04239-4_33
14. Luo, X., et al.: SCPM-net: An anchor-free 3D lung nodule detection network using sphere representation and center points matching. arXiv preprint arXiv:2104.05215 (2021)
15. Nguyen, E.H., et al.: Circle representation for medical object detection. IEEE Transactions on Medical Imaging (2021)
16. Peng, S., Jiang, W., Pi, H., Li, X., Bao, H., Zhou, X.: Deep snake for real-time instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8533–8542 (2020)
17. Puelles, V.G., Hoy, W.E., Hughson, M.D., Diouf, B., Douglas-Denton, R.N., Bertram, J.F.: Glomerular number and size variability and risk for kidney disease. Curr. Opin. Nephrol. Hypertens. **20**(1), 7–15 (2011)
18. Rehem, J.M.C., Dos Santos, W.L.C., Duarte, A.A., De Oliveira, L.R., Angelo, M.F.: Automatic glomerulus detection in renal histological images. In: Medical Imaging 2021: Digital Pathology, vol. 11603, p. 116030K. International Society for Optics and Photonics (2021)
19. Yang, H., et al.: Circle Net: Anchor-Free Glomerulus Detection with Circle Representation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 35–44. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_4
20. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2403–2412 (2018)
21. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)



Vertebrae Localization, Segmentation and Identification Using a Graph Optimization and an Anatomic Consistency Cycle

Di Meng(✉), Eslam Mohammed, Edmond Boyer, and Sergi Pujades

CNRS, Grenoble INP, LJK, Inria, University Grenoble Alpes, Grenoble, France
di.meng@inria.fr

Abstract. Vertebrae localization, segmentation and identification in CT images is key to numerous clinical applications. While deep learning strategies have brought to this field significant improvements over recent years, transitional and pathological vertebrae are still plaguing most existing approaches as a consequence of their poor representation in training datasets. Alternatively, proposed non-learning based methods take benefit of prior knowledge to handle such particular cases. In this work we propose to combine both strategies. To this purpose we introduce an iterative cycle in which individual vertebrae are recursively localized, segmented and identified using deep networks, while anatomic consistency is enforced using statistical priors. In this strategy, the transitional vertebrae identification is handled by encoding their configurations in a graphical model that aggregates local deep-network predictions into an anatomically consistent final result. Our approach achieves the state-of-the-art results on the VerSe20 challenge benchmark, and outperforms all methods on transitional vertebrae as well as the generalization to the VerSe19 challenge benchmark. Furthermore, our method can detect and report inconsistent spine regions that do not satisfy the anatomic consistency priors. The code and model are available for research purposes. (https://gitlab.inria.fr/spine/vertebrae_segmentation)

Keywords: Spine · Vertebrae segmentation · Vertebrae labelling · Transitional vertebrae

1 Introduction

Segmenting, identifying and localizing vertebrae in CT images play an essential role in many clinical applications [16]. These tasks are intrinsically interdependent and a common issue comes from inconsistencies that can propagate between them. For instance neighboring vertebrae share similar shapes which

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_32.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 307–317, 2022.
https://doi.org/10.1007/978-3-031-21014-3_32

makes their identification uncertain. Moreover pathological spines can present abnormal shapes or the number of transitional vertebrae can be different among patients. These transitional vertebrae, *i.e.*, the absence of T12 or occurrence of T13 or L6, are common and are reported to affect between 15% and 35% of the general population [32]. However, they only impact one vertebrae among the 24 in a spine and their occurrence percentages in standard vertebrae datasets appear to be therefore much lower, down to a few percentages in practice. As a consequence, the current state-of-the-art methods, which are mostly based on deep networks trained over standard datasets, tend to experience important performance drops in the presence of transitional vertebrae.

Besides augmenting datasets, an alternative solution is to exploit prior knowledge on the full structure of the spine, as initially introduced in non-learning based methods. In this work, we investigate the combination of such a strategy with a deep learning approach. Precisely we propose to iteratively cycle between the localization, segmentation and identification tasks, which uses deep networks, while enforcing anatomical consistency with statistical priors. On one hand, the priors are used to localize vertebrae: they leverage learned statistics of vertebrae volume and inter-vertebral distances which exhibit more robustness to pathological cases than shape or appearance models. On the other hand, in order to handle transitional vertebrae in the identification, we propose to encode the admissible configurations in a graphical model. Such a model leverages local deep-network predictions and aggregates them into an anatomically consistent result. Our experiments demonstrate that this strategy successfully handles local inaccurate predictions. Therefore it performs better than other methods with transitional vertebrae while providing state-of-the-art results on standard benchmarks such as VerSe20.

To solve for vertebrae localization, segmentation and identification, traditional methods use shape [5] or graphical models [4,10]. Surface [2,5,15,23] or image based appearance models [4,10], which make assumptions on exact shapes, are not robust to pathological cases (such as fractures or metal implants). However relative positions of vertebrae [10,11,33] are less sensitive to these cases. In our work we propose to learn statistics on vertebrae relative positions and volumes. The volume quantifies a *part of the vertebra shape* that is robust to morphological deviations. Furthermore we learn statistics conditioned on the spine level (cervical, thoracic, lumbar), increasing hence their performance. Moreover we relax the need for a fixed threshold to detect vertebrae [25] and compute adaptative thresholds that automatically vary across spine levels and patients.

For the segmentation task, existing methods use primarily statistical models [15,26], active shape models [1,3,12], graph cuts [2], or level sets [31]. Several methods achieve very good performances by assuming that a specific part of the spine is observed [14,26], yet making them sensitive to arbitrary field of views. For this task, recent learning based approaches have demonstrated the best performances [17,21,25] with a local approach. Our segmentation strategy builds on the segmentation network of Payer *et al.* [25] and the refinement strategy of Lessman *et al.* [17].

As for the vertebrae identification task, fully automatic methods rely on machine learning and convolutional neural networks (CNN) in which hand-crafted

features [7,11] tend to be replaced by learned ones. Several methods [7,13,22,25] use a two-step approach that first detects the vertebrae before identifying them. Other networks, such as Btrfly [28], bidirectional recurrent network (Bi-RNN) [18], FCN [8] or the Spine-transformer [30] output directly vertebrae locations and labels simultaneously. All these methods have difficulties with transitional vertebrae as a consequence of their sparsity in training data. Another group of methods proposes to enforce the identification consistency as a post process by using chain-structure graphical models [34], conditional random fields [20] or Hidden Markov model [8]. They anyway do not explicitly model transitional vertebrae. In our work, we take a two-step approach to localize and identify vertebrae. To identify individual vertebra we build on Mohammed *et al.* [24] and use the segmented vertebrae shape instead of the raw CT image. In addition we also contribute a new graphical model to select a global optimal configuration among all the individual predictions, within which the transitional vertebrae are explicitly modeled.

2 Method

2.1 Anatomic Consistency Cycle

We propose to cycle through the three tasks of localization, segmentation and identification in order to enforce anatomic consistencies among them. Our method starts by segmenting a *spine mask* (details in Sup Mat.) which enables to locate individual vertebrae (Sect. 2.2). From the obtained locations, individual vertebrae segmentation masks are estimated with an individual vertebrae segmentor [25] and refined with an iterative location-segmentation refinement scheme [17]. Vertebrae are further identified using the locations and segmentation masks (Sect. 2.3). The obtained identifications are in turn used to enforce finer anatomic consistency constraints (Sect. 2.2) and detect possible new candidate locations. In the latter case, the new vertebrae go through the segmentation and identification steps described before. The process ends when the proposed consistency criteria are satisfied. It also stops when the set of detected locations, segmentation masks and identifications does not change over the cycle. The remaining inconsistencies, such as a failure in the location-segmentation refinement or an inconsistency in the anatomic constraints, are reported in the result. Figure 1 illustrates the method overview.

2.2 Anatomic Consistency Constraints for Vertebrae Localization

Vertebrae are naturally ordered in the spine and their relative locations and consecutive sizes are heavily correlated. With the objective to exploit such prior information we compute statistics over inter-vertebral distances and individual vertebrae volumes. Given these priors we enforce two consistency criteria. First, distances between all detected locations should follow the inter-vertebral distances statistics. Second, the *spine mask* should be similar to the union of individual vertebrae masks. The segmentation of a *spine mask* is, in general, a straightforward task which in turn allows the areas where individual vertebrae

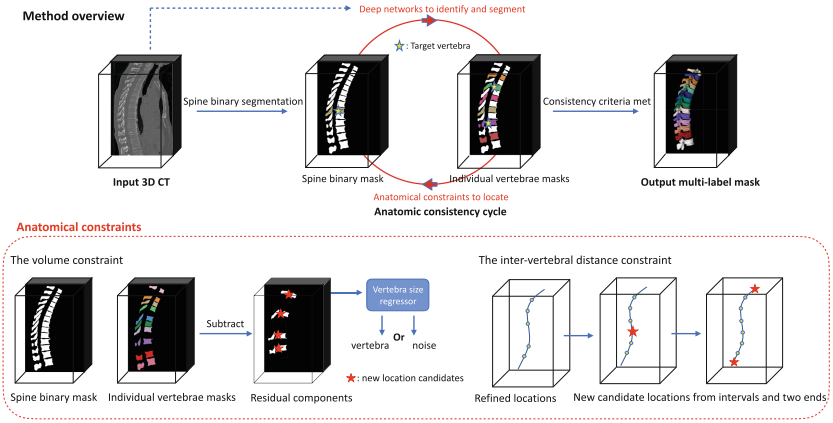


Fig. 1. The method overview. Given a 3D CT as input, the spine is segmented as a reference to locate the individual vertebrae. The anatomical constraints are leveraged with deep networks for localization and segmentation. Once the location is stable, its identification is obtained given its segmentation. The set of location, segmentation and identification is cycled through until the consistency criteria are met.

should lie to be identified. In practice, at each cycle the difference between the *spine mask* and all individual vertebrae masks is computed to obtain residual connected components. (The residual connected component is the whole *spine mask* at the first cycle.) Vertebrae volume statistics are then considered to decide whether a residual is a vertebra or just noise.

The Vertebrae Volume Constraint. Given a residual connected component, the idea is to check its volume by considering the volumes of the neighboring vertebrae. To this purpose a regressor is trained over consecutive vertebrae to predict the volume of the previous or next vertebrae along the spine. If the residual volume size is larger than a fraction of the predicted size (50% in our experiments), it is considered as a vertebra and its location at its mass center is added to the vertebrae list. The residual is otherwise discarded as noise.

The Inter-Vertebral Distance Constraint. This constraint builds on the fact that distances between locations are well structured. To this aim, we use two statistical models of the distance between vertebrae. The first is a Gaussian distribution for each anatomic group capturing the mean and variance of the distances between consecutive vertebrae. It is used as a prior to detect abnormally large distances. In addition, linear regressors conditioned on each anatomic group are used to predict the inter-vertebral distance from the neighbouring ones. When a larger inter-vertebral distance is detected, new candidate locations are added in between. The number of new candidates is adapted depending on the current distance and the predicted one. We also leverage identification to check if the spine extremes are complete. If C1 (or L5) is not yet found, a location is predicted up (or down) using the two most top (down) locations. If it is inside the image field of view, it is added.

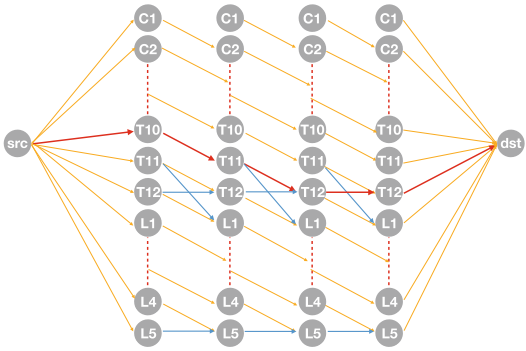


Fig. 2. Spine identification global graph. An example with 4 vertebrae. Edges in orange are regular connections between nodes. Edges in blue connecting T12 to T12, and L5 to L5, allow the presence of T13 and L6. The edge between T11 and L1 allows T12 to be absent. The result with red edges corresponds to a spine where T10, T11, T12 and T13 are observed. (Best viewed in color)

2.3 Vertebrae Identification

To identify vertebrae we propose to combine both a local and global reasoning. The local reasoning is based on an individual classification network that predicts one of the 24 possible labels for each vertebra. On top of it, the global reasoning assumes that consecutive vertebrae are *sorted* in the spine and aggregates the local predictions into a global, consistent identification.

Individual Vertebra Classification. Our individual vertebra classification module builds on [24] where a 3D VGG network is used to classify a single vertebra into one of the 24 categories. The input of the network is a binary volume of size $128 \times 128 \times 128$ that contains a vertebra segmentation mask. In our work, we extend this method [24] by including the neighboring vertebrae to the input, while keeping the target vertebra in the center of the cube, which effectively adds more context for the prediction. At inference time, two predictions are performed: one with the *spine mask* and a second with the union of all individual vertebrae segmentations. Both predictions are averaged and used as input to the graph optimization. When an individual segmentation mask is empty, *e.g.* metal implants can disturb the individual segmentor, a uniform distribution over all vertebrae is used. In order to deal with transitional vertebrae, such as T13 and L6, for which only few samples appear in the training set - 1 T13 and 13 L6 in practice - we consider T13 as T12 and L6 as L5 and leave their fine identification to the global reasoning.

Global Graph Optimization. Some vertebrae can still be misidentified by the local approach. In order to improve the classification, we propose a global strategy that enforces the natural ordering of vertebrae, i.e. consecutive vertebrae should have consecutive labels. We cast the problem as a shortest path search in a graph which identifies an optimal set of labels, as illustrated in Fig. 2.

Given a list with n consecutive locations, a graph with $n \times 24$ nodes is created. Each of the n columns represents a located vertebra, and each of the 24 rows represents a vertebra class. Both a unary and a binary costs are then considered for the optimal path search. We populate the unary cost of each node with the cost obtained from the individual local probability. Additionally, as the 3-group predictions are robust, we also add a cost to each unary to prevent group swapping. The binary costs are defined on the edges between the nodes in the graph. These edges encode the fact that two consecutive vertebrae need to have consistent classes. In practice, node N_i^j with $i \in [1, n]$ and $j \in [1, 24]$ is only connected with N_{i+1}^{j+1}. These edges are shown in orange in Fig. 2. They effectively enforce that consecutive vertebrae get consecutive labels.

Three configurations require special attention in this process, which correspond to transitional vertebrae: i) The presence of T13, ii) The lack of T12 and iii) The presence of L6. To handle the presence of T13, an edge is added between two consecutive T12 nodes. For the lack of T12, an edge is added between two consecutive nodes belonging to T11 and L1. The presence of L6 is dealt with an edge between two consecutive L5 nodes. These three special edges are shown in blue in Fig. 2 and are given a higher cost. To complete the graph, two extra nodes with no cost are added that are used as the source (*src*) and the destination (*dst*) in the optimization. The *src* node is connected to all nodes of the first vertebrae and the *dst* node is connected to all nodes of the last vertebrae.

Once the graph is created and the costs are populated, we compute the shortest path in the graph using the classical Dijsktra algorithm [9]. As a post process, we check if repeated instances of T12 (or L5) are obtained and adjust the class of the second node to T13 (or L6) accordingly.

3 Experiments

We evaluate our method using the VerSe20 MICCAI challenge dataset [19,27,29], which is currently the largest spine dataset. It is split into a public training set, a public testset and a hidden testset. Each set has 100 CT scans. The challenge has closed so the annotation of the hidden testset is also released for the public. We follow the challenge setting and use the training data to set up our method: 80 random CT scans from the training set are used to train the networks and estimate the anatomical prior statistics, the 20 left for validation. We first evaluate the novel graphical model performance and then we quantitatively evaluate our full method on the public and hidden testsets.

Vertebrae Identification. To evaluate the benefit of using the context and the graph (Sect. 2.3), we compare four identification settings: Using a single vertebra binary mask [24] (a), adding the neighbouring vertebrae (b) and adding the neighbouring vertebrae and using the graph (c). To compare to image based identification approaches we trained a classification method using the CT image as input (d). For a) the identification accuracy is 70.50%. When including the neighboring vertebrae masks, the accuracy is increased to 84.55%. After the graph optimization, 97.36% accuracy is obtained. When using the CT image

Table 1. Quantitative comparison of our method with the top ranked methods on the VerSe20 challenge [27] as well as with alternative (a) in which the anatomic consistency constraints are not conditioned on the identification and alternative (b) where the anatomic consistency priors are not used for detection.

	Public testset				Hidden testset			
	Labelling results		Segmentation results		Labelling results		Segmentation results	
	ID rate (%)	MLD (mm)	DSC (%)	HD (mm)	ID rate (%)	MLD (mm)	DSC (%)	HD (mm)
Zhang A	94.93	2.99	88.82	7.62	96.22	2.59	89.36	7.92
Payer C. [25]	95.06	2.90	91.65	**5.80**	92.82	2.91	89.71	**6.06**
Chen D. [6]	95.61	**1.98**	91.72	6.14	**96.58**	**1.38**	**91.23**	7.15
Ours	**96.59**	2.21	**92.53**	7.03	95.38	2.43	91.11	6.69
Ours(a)	96.02	2.22	92.06	6.82	94.93	2.34	90.85	6.53
Ours(b)	93.61	3.10	89.76	7.80	89.41	3.81	85.45	8.39

patch as input instead of its mask, a lower identification accuracy (82.36%) is obtained compared to the binary mask of the vertebrae (84.55%). We believe that this is due to the fact that the masks provide shape discriminative cues on the vertebrae.

VerSe20 Challenge Results. We quantitatively evaluate our method on the VerSe20 Challenge public and hidden testsets with 200 CT subjects in total and adopted the metrics of the challenge evaluation protocol [27]: The *Dice coefficient (DSC)* and the *Hausdorff Distance (HD)* evaluate the correctly identified segmentation performance in terms of voxel and surface similarity respectively; The *Identification rate (ID rate)* and the *Mean Localization Distance (MLD)* evaluate the accuracy of the labelling task, where MLD measures the Euclidean distance of the predicted location to the GT location and 20mm is the criterion for a valid identification. In the challenge metrics, when a vertebra is missed, MLD and HD are not computed. Thus there is a trade off between ID rate and MLD (same for DSC and HD). While more positive detections lead to a better ID rate, a worse MLD can be obtained if the detections are less accurate.

Table 1 shows the results of the proposed method and the top scoring methods. Our method is on pair with the best performing method in the leaderboard [6] while being significantly more robust to transitional cases, as shown in Table 2, which is our main objective. In that respect, we observe that Chen D. et al. [6] did not win the challenge, as one important criterion in the challenge was the performance in handling the transitional vertebrae - 6 cases with T13 (2/2/2 in Train/Public/Hidden), 47 cases with L6 (15/15/17) and 8 cases with absent T12(3/4/1). Following the challenge guidelines we computed the Dice score only on scans with transitional vertebrae and the obtained results are presented in Table 2. Our method consistently outperforms all existing methods.

To evaluate the generalization of our method, we perform the experiment of the Verse20 challenge [27] where the approach developed and trained on the VerSe20 training set is applied to the VerSe19 hidden testset. Table 3 shows that our method outperforms all approaches and generalizes well to VerSe19.

Table 2. Methods evaluation (Dice score %) on transitional vertebrae [27].

	Public testset	Hidden testset
Ours	**91.04**	**89.70**
Payer C. [25]	85.96	89.59
Zhang A	87.15	87.35
Chen D. [6]	84.21	87.01

Table 3. Generalization performance on VerSe19 hidden testset [27].

	DSC(%)
Ours	**90.84**
Chen D. [6]	86.44
Payer C. [25]	84.11
Zhang A	85.42

Inference times on public testset vary between 1.5 mins to 248 mins depending on the scan size, with a median runtime of 26 mins. Most of the computation time is devoted to the spine segmentation, which could be drastically improved by considering parallelization over patches.

Qualitative results and typical failure cases of our method are provided in the supplementary material.

Ablation Study. To highlight the benefits of the different parts of the proposed method we perform two experiments. In the first one (*ours(a)* in Table 1) we do not use the identification to condition the vertebrae statistics. Instead of learning specialized regressors for each spine level, we use a fixed threshold to determine if a residual mask is a potential true vertebra (we use half of smallest vertebra volume size in the training set: 7820/2). To decide if there is a missing vertebra given a distance between 2 we use a fixed threshold of 50 mm [25]. The overall scores show that the improvement when using conditional statistics is minor but not negligible: without conditioning 8 vertebrae are missed in the public testset and 6 in the hidden test set while with the identification information no vertebra is missed in public testset and only 4 in hidden testset. In the second experiment (*ours(b)* in Table 1) the anatomic consistency priors (Sect. 2.2) are not used. Without the constraints on vertebrae volume and the relative locations the performance significantly drops.

4 Conclusion

In this work we presented a new approach to localize, segment and identify vertebrae in CT images. An anatomic consistency cycle is proposed that aggregates task-specific deep networks while enforcing statistical priors. As seen with the generalization on the VerSe19 dataset, it improves robustness in the results. In addition, we also present a graphical model that enforces global consistency among individual predictions, hence being able to handle transitional vertebrae as well as pathological cases. The evaluation on the standard VerSe20 challenge benchmark demonstrates the interest of the proposed strategy, in particular with transitional vertebrae which are present in a meaningful part of the population. The extension of the proposed anatomic consistency cycle to other anatomic structures with similar specific cases is currently under investigation.

References

1. Al Arif, S.M.R., Gundry, M., Knapp, K., Slabaugh, G.: Improving an active shape model with random classification forest for segmentation of cervical vertebrae. In: Yao, J., Vrtovec, T., Zheng, G., Frangi, A., Glocker, B., Li, S. (eds.) CSI 2016. LNCS, vol. 10182, pp. 3–15. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-55050-3_1
2. Aslan, M.S., Ali, A., Chen, D., Arnold, B., Farag, A.A., Xiang, P.: 3d vertebrae segmentation using graph cuts with shape prior constraints. In: ICIP. IEEE (2010)
3. Benjelloun, M., Mahmoudi, S., Lecron, F.: A framework of vertebra segmentation using the active shape model-based approach. IJBI (2011)
4. Bromiley, P.A., Kariki, E.P., Adams, J.E., Cootes, T.F.: Fully automatic localisation of vertebrae in ct images using random forest regression voting. In: Yao, J., Vrtovec, T., Zheng, G., Frangi, A., Glocker, B., Li, S. (eds.) CSI 2016. LNCS, vol. 10182, pp. 51–63. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-55050-3_5
5. Cai, Y., Osman, S., Sharma, M., Landis, M., Li, S.: Multi-modality vertebra recognition in arbitrary views using 3d deformable hierarchical model. TMI (2015)
6. Chen, D., Bai, Y., Zhao, W., Ament, S., Gregoire, J., Gomes, C.: Deep reasoning networks for unsupervised pattern de-mixing with constraint reasoning. In: ICML. PMLR (2020)
7. Chen, H., et al: Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 515–522. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_63
8. Chen, Y., Gao, Y., Li, K., Zhao, L., Zhao, J.: vertebrae identification and localization utilizing fully convolutional networks and a hidden markov model. TMI (2019)
9. Dijkstra, E.W., et al.: A note on two problems in connexion with graphs. Numerische mathematik (1959)
10. Glocker, B., Feulner, J., Criminisi, A., Haynor, D.R., Konukoglu, E.: Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7512, pp. 590–598. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33454-2_73
11. Glocker, B., Zikic, D., Konukoglu, E., Haynor, D.R., Criminisi, A.: Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8150, pp. 262–270. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40763-5_33
12. Graham, J., Cooper, D., Taylor, C., Cootes, T.: Active shape models their training and applications. CVIU (1995)
13. Jakubicek, R., Chmelik, J., Jan, J., Ourednicek, P., Lambert, L., Gavelli, G.: Learning-based vertebra localization and labeling in 3d CT data of possibly incomplete and pathological spines. Comput. Methods Programs Biomed **123** (2020)
14. Janssens, R., Zeng, G., Zheng, G.: Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3d fully convolutional networks. In: ISBI (2018)
15. Klinder, T., Ostermann, J., Ehm, M., Franz, A., Kneser, R., Lorenz, C.: Automated model-based vertebra detection, identification, and segmentation in CT images. MedIA (2009)

16. Knez, D., Likar, B., Pernuš, F., Vrtovec, T.: Computer-assisted screw size and insertion trajectory planning for pedicle screw placement surgery. TMI (2016)
17. Lessmann, N., Van Ginneken, B., De Jong, P.A., Išgum, I.: Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. MedIA (2019)
18. Liao, H., Mesfin, A., Luo, J.: Joint vertebrae identification and localization in spinal CT images by combining short-and long-range contextual information. TMI (2018)
19. Löffler, M.T., et al.: A vertebral segmentation dataset with fracture grading. Radiol. Artif. Intell. **2** (2020)
20. Mader, A.O., Lorenz, C., von Berg, J., Meyer, C.: Automatically localizing a large set of spatially correlated key points: a case study in spine imaging. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 384–392. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_43
21. Masuzawa, N., Kitamura, Y., Nakamura, K., Iizuka, S., Simo-Serra, E.: Automatic segmentation, localization, and identification of vertebrae in 3D CT images using cascaded convolutional neural networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 681–690. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_66
22. McCouat, J., Glocker, B.: Vertebrae detection and localization in CT with two-stage CNNS and dense annotations. arXiv preprint arXiv:1910.05911 (2019)
23. Meng, D., Keller, M., Boyer, E., Black, M., Pujades, S.: Learning a statistical full spine model from partial observations. In: Reuter, M., Wachinger, C., Lombaert, H., Paniagua, B., Goksel, O., Rekik, I. (eds.) ShapeMI 2020. LNCS, vol. 12474, pp. 122–133. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61056-2_10
24. Mohammed, E., Meng, D., Pujades, S.: Morphology-based individual vertebrae classification. In: Reuter, M., Wachinger, C., Lombaert, H., Paniagua, B., Goksel, O., Rekik, I. (eds.) ShapeMI 2020. LNCS, vol. 12474, pp. 134–144. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61056-2_11
25. Payer, C., Stern, D., Bischof, H., Urschler, M.: Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and u-net. In: VISIGRAPP (5: VISAPP) (2020)
26. Rasoulian, A., Rohling, R., Abolmaesumi, P.: Lumbar spine segmentation using a statistical multi-vertebrae anatomical shape+ pose model. TMI (2013)
27. Sekuboyina, A., et al.: Verse: a vertebrae labelling and segmentation benchmark for multi-detector CT images. MedIA (2021)
28. Sekuboyina, A., et al.: Btrfly net: vertebrae labelling with energy-based adversarial learning of local spine prior. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 649–657. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_74
29. Sekuboyina, A., Rempfler, M., Valentinitsch, A., Menze, B.H., Kirschke, J.S.: Labeling vertebrae with two-dimensional reformations of multidetector CT images: an adversarial approach for incorporating prior knowledge of spine anatomy. Radiol. Artif. Intell. **2** (2020)
30. Tao, R., Liu, W., Zheng, G.: Spine-transformers: vertebra labeling and segmentation in arbitrary field-of-view spine CTS via 3d transformers. MedIA (2022)
31. Tsai, A., et al.: A shape-based approach to the segmentation of medical imagery using level sets. TMI (2003)
32. Uçar, D., et al.: Retrospective cohort study of the prevalence of lumbosacral transitional vertebra in a wide and well-represented population. Arthritis (2013)

33. Wang, F., Zheng, K., Lu, L., Xiao, J., Wu, M., Miao, S.: Automatic vertebra localization and identification in CT by spine rectification and anatomically-constrained optimization. In: Proceedings of the IEEE/CVF CVPR (2021)
34. Yang, D., et al.: Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 633–644. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_50



Coronary Ostia Localization Using Residual U-Net with Heatmap Matching and 3D DSNT

Milosz Gajowczyk[1,2], Patryk Rygiel[1,2], Piotr Grodek[1], Adrian Korbecki[1,3], Michal Sobanski[1,3], Przemyslaw Podgorski[1,3], and Tomasz Konopczynski[1(✉)]

[1] Hemolens Diagnostics Ltd., Wroclaw, Poland
tomasz.konopczynski@hemolens.eu
[2] Wroclaw University of Science and Technology, Wroclaw, Poland
[3] Wroclaw Medical University, Wroclaw, Poland

Abstract. Localization of coronary ostia landmarks in Computed Tomography Angiography (CTA) volumes is a crucial step in developing various automatic diagnostic procedures. In this study, we propose a one-step method of coronary ostia landmark localization that utilizes a residual U-Net with heatmap matching and 3D Differentiable Spatial to Numerical Transform (DSNT). We evaluate the method using two datasets: a Coronary Computed Tomography Angiography (CCTA) dataset containing 201 scans and a publicly available ImageTBAD dataset containing 77 CTA scans annotated with coronary ostia landmarks.

On the CCTA dataset we report median Euclidean distance error – 1.14 mm on the left coronary ostium and 0.98 mm on the right coronary ostium. On the ImageTBAD CTA dataset we report median Euclidean distance error – 3.48 mm on the left coronary ostium and 2.97 mm on the right coronary ostium. Our evaluation shows that the proposed method improves accuracy of coronary ostia landmark localization when compared to other known methods.

Keywords: Artificial neural networks · Deep learning · Landmark regression · Coronary ostia localization · Computed tomography

1 Introduction

An automatic coronary ostia localization is used in various medical image analysis techniques. It is often paired with other tasks such as an image segmentation or a registration. The main challenges in automatic coronary ostia localization are anatomical differences between patients and differences in image acquisition.

The first methods which tackle the problem of coronary ostia localization from a CTA or CCTA are based on finding intensity patterns. Both Waechter et al. [26] and Elattar et al. [5] use a prior aortic valve segmentation to perform

M. Gajowczyk and P. Rygiel—Equal contribution.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 318–327, 2022.
https://doi.org/10.1007/978-3-031-21014-3_33

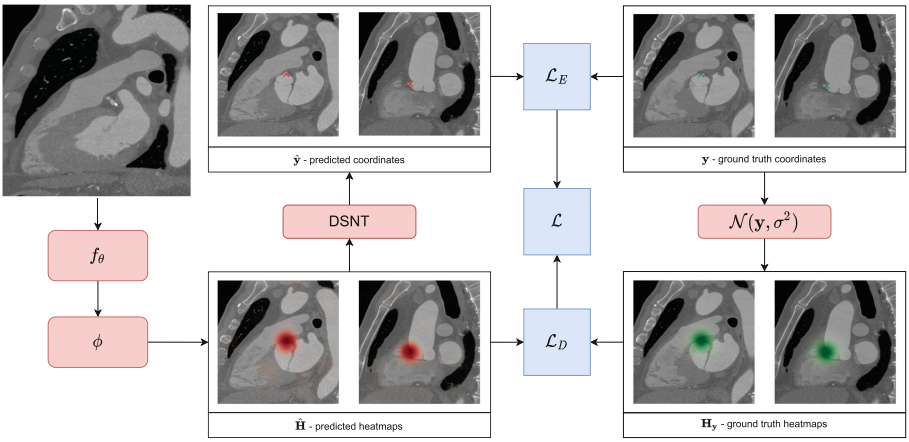


Fig. 1. Diagram of our method with an example from the CCTA dataset. The input volume is processed by a residual U-Net f_θ and a normalization function ϕ to predict heatmaps $\widehat{\mathbf{H}}$. These heatmaps are then transformed by a DSNT layer to coordinates predictions $\widehat{\mathbf{y}}$. The loss $\mathcal{L}$ consists of the Euclidean distance $\mathcal{L}_E$ between predicted coordinates $\widehat{\mathbf{y}}$ and ground-truth coordinates $\mathbf{y}$, and a divergence $\mathcal{L}_D$ between predicted heatmaps $\widehat{\mathbf{H}}$ and ground-truth heatmaps $\mathbf{H}_\mathbf{y}$. Ground-truth heatmaps $\mathbf{H}_\mathbf{y}$ are modeled as a Gaussian distribution $\mathcal{N}(\mathbf{y}, \sigma^2)$. Predicted and ground-truth localization and heatmaps are highlighted in red and green, respectively.

an intensity-based landmark localization on its surface. Elattar et al. report an Euclidean error of 2.81 ± 2.07 mm.

The introduction of machine learning was beneficial both for robustness and for time consumption. Classical machine learning methods can be divided into two types: a classification-based [3,4,8,12,13,19,25,31] and a regression-based [1,4,6,7,19,22,25] methods. The classification-based methods are producing a probability for a given pixel or voxel to be a landmark of interest. For the application in the coronary ostia localization Ionasec et al. [8] propose to use probabilistic boosting trees and report an Euclidean error of 2.28 mm. Zheng et al. [31] use global-to-local marginal space learning and report an Euclidean error of 2.11 ± 1.34 mm.

The regression-based methods are designed to output a continuous value that defines sub-resolution coordinates for a landmark of interest. Al et al. [1] employ an approach based on colonial walks to the problem of coronary ostia localization and report an Euclidean error of 2.01 ± 1.04 mm. The first Deep Neural Networks (DNN) which tackle the problem of landmark detection employ a Convolutional Neural Network (CNN) model with a fully-connected head to regress landmark coordinates [9,24]. Mostafa et al. [14] propose to use a U-Net architecture [21] with a fully-connected head to solve the problem of coronary ostia localization. They report 68.3% and 74.1% SDR (success detection rate) for left and right coronary ostium localization, respectively, for the threshold of 2 mm.

In subsequent work, the idea of using heatmap matching is introduced [23]. In this approach, a landmark is represented as a Gaussian centered at its coordinates. This landmark representation is similar to classification-based machine learning methods where a probability of a landmark presence at its position is predicted for each pixel or voxel. The popular landmark localization models utilize the pyramid or hourglass architecture [15,28]. Heatmap matching is also used for medical images by Payer et al. [20] to perform a landmark localization on X-ray images. Wolterink et al. [27] use it in coronary ostia localization and obtain an Euclidean error of 1.8 ± 1.0 mm for a test set of 36 images.

Noothout et al. developed in 2018 an approach that performed a slice-wise image analysis [18]. This approach comprises a vector pointing from the center of each slice to a landmark of interest. They report the Euclidean error of 2.88 and 2.19 mm for the left and right coronary ostium, respectively, for a test set of 40 CCTA volumes. Noothout's approach is further refined by its creators and expanded in 2020 by Noothout et al. [17]. The slice-wise approach is changed to the patch-wise approach and a CCTA dataset is further expanded to 672 volumes with 200 volumes in a test set. With a larger dataset and a greater anatomical variability, they achieve a median Euclidean error of 1.45 and 1.55 mm for the left and the right coronary ostium, respectively and interquartile range (IQR) of 0.97 and 1.20 mm for the left and the right coronary ostium, respectively. This is the highest reported accuracy of coronary ostia localization known to the authors. However, such results are only achieved by using a subsequent local refinement network. The lowest median (IQR) Euclidean error achieved with the use of a global network only is reported to be 3.31 (2.23), 2.90 (1.93) mm for the left and the right coronary ostium, respectively.

In this work, we propose a one-stage approach to the problem of 3D coronary ostium landmark localization in CT volumes. Our approach combines heatmap matching regression with a 3D differentiable spatial to numerical transform (DSNT) [16]. Regression of landmarks' coordinates is done globally and does not require local refinement as other most common techniques [5,17,18,26,31]. We evaluate the method using two datasets: an in-house CCTA dataset and a publicly available ImageTBAD CTA dataset. It should be noted, that the clarity of ostia location in CTA in ImageTBAD is significantly lower compared to the CCTA dataset and therefore more challenging. ImageTBAD dataset contains volumes of lower resolution with many artefacts and consist of aortic non-ECG-gated CTA volumes which are not dedicated for coronary ostia imaging.

2 Method

In this work, we consider the problem of learning a supervised regression model of coronary ostia coordinates from a CT volume and labels acquired from human annotations. We refer to input volume as $\mathbf{x}$, human annotations as $\mathbf{y}$ and model predictions as $\widehat{\mathbf{y}}$. Our proposed approach combines heatmap matching regression with extraction of landmark coordinates from a heatmap using a DSNT layer. We describe details of the method in the following sections and the sketch of the method is presented in Fig. 1.

2.1 Heatmap Generation

In general, the unnormalized heatmap $\boldsymbol{H} = f_\Theta(\mathbf{x})$ is an output from a model f parameterized by learnable parameters Θ with input volume $\mathbf{x}$. The model f_Θ that acts as a backbone in this work is the residual U-Net [10,30] that adopts a classic U-Net [21] architecture and combines it with a residual neural network. Each residual unit can be defined as follows:

$$\begin{aligned} \mathbf{y}_l &= h(\mathbf{x}_l) + \mathcal{F}(\mathbf{x}_l, \mathbf{W}_l), \\ \mathbf{x}_{l+1} &= f(\mathbf{y}_l), \end{aligned} \tag{1}$$

where $\mathbf{x}_l$ and $\mathbf{x}_{l+1}$ are the input and output of the l-th residual unit, $\mathcal{F}(\cdot)$ is the residual function, $f(\mathbf{y}_l)$ is an activation function and $h(\mathbf{x}_l)$ is an identity mapping function. The identity mapping function is usually of a form $h(\mathbf{x}_l) = \mathbf{x}_l$.

2.2 DSNT

The Differentiable Spatial to Numerical Transform (DSNT) [16] is a fully differentiable and non-trainable layer. Due to this fact, numerical coordinates can be obtained during training and can be used in a learning criterion since backpropagation does pass via DSNT. We define the 3D extension of the DSNT as follows:

$$\mathrm{DSNT}(\widehat{\boldsymbol{H}}) = \Big[\langle \widehat{\boldsymbol{H}},\ \boldsymbol{X} \rangle_F \quad \langle \widehat{\boldsymbol{H}},\ \boldsymbol{Y} \rangle_F \quad \langle \widehat{\boldsymbol{H}},\ \boldsymbol{Z} \rangle_F \Big], \tag{2}$$

where $\widehat{\boldsymbol{H}} = \phi(\boldsymbol{H})$ is a normalized heatmap $\boldsymbol{H}$ output from the backbone $f_\Theta(\mathbf{x})$ by a normalization function $\phi(\cdot)$, $\langle \cdot, \cdot \rangle_F$ represents Frobenius inner product and $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are grids in a form of $n \times m \times l$ matrices used to calculate expected values of landmarks' coordinate and are defined as:

$$\boldsymbol{X}_{i,\ j,\ k} = \frac{2j - (n+1)}{n}, \quad \boldsymbol{Y}_{i,\ j,\ k} = \frac{2i - (m+1)}{m}, \quad \boldsymbol{Z}_{i,\ j,\ k} = \frac{2k - (l+1)}{l}, \tag{3}$$

that is, each entry of $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ contains its own coordinates, scaled such that the top-left-upper corner of the volume is at $(-1, -1, -1)$ and bottom-right-down is at $(1, 1, 1)$.

2.3 Loss Functions

During training we utilize the popular Euclidean metric $\mathcal{L}_E$ between the network output $\widehat{\mathbf{y}} = \mathrm{DSNT}(\widehat{\mathbf{H}})$ and the ground-truth $\mathbf{y}$. It has been shown [16] that forcing the network to produce gaussian-like heatmaps improves the overall prediction. To achieve that we use the Jensen-Shannon divergence regularization $\mathcal{L}_D$ as a term to calculate, and minimize the divergence between heatmap prediction $\widehat{\mathbf{H}}$ and appropriate target normal distribution $\mathcal{N}(\mathbf{y}, \sigma^2)$ with σ being a hyper-parameter controlling the spread of the desired distribution.

The complete loss is defined as follows:

$$\mathcal{L}(\widehat{\mathbf{H}},\ \mathbf{y}) = \mathcal{L}_E\big(\mathrm{DSNT}(\widehat{\mathbf{H}}),\ \mathbf{y}\big) + \lambda \mathcal{L}_D(\widehat{\mathbf{H}},\ \mathbf{y}), \tag{4}$$

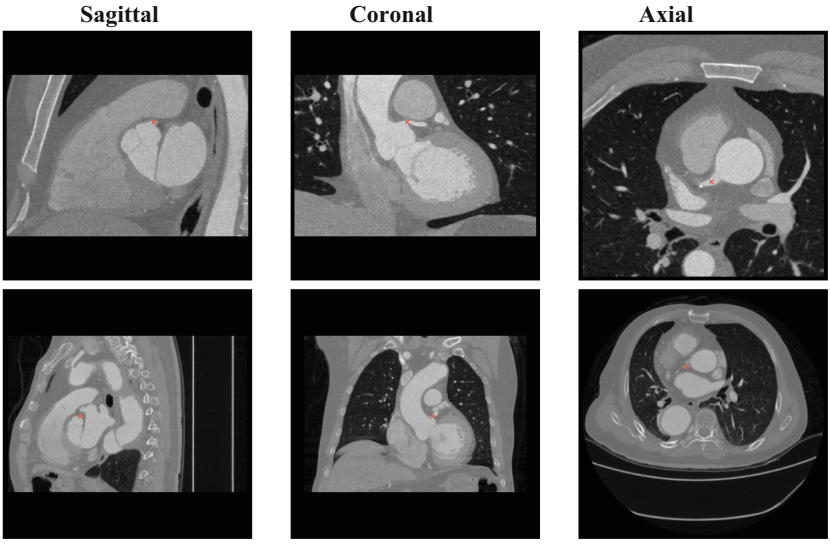


Fig. 2. Example slices from datasets with ostium annotation highlighted in red. The first row shows examples from CCTA, the second from ImageTBAD dataset. (Color figure online)

with λ being a regularization coefficient hyper-parameter controlling the strength of the regularization. The Euclidean loss $\mathcal{L}_E$ is defined as:

$$\mathcal{L}_E(\widehat{\mathbf{y}}, \mathbf{y}) = \|\widehat{\mathbf{y}} - \mathbf{y}\|_2, \tag{5}$$

and the Jensen-Shannon divergence regularization $\mathcal{L}_D$ is

$$\mathcal{L}_D(\widehat{\mathbf{H}}, \mathbf{y}) = D_{JS}(\widehat{\mathbf{H}} \parallel \mathcal{N}(\mathbf{y}, \sigma^2)), \tag{6}$$

where $D_{JS}(\cdot||\cdot)$ is the Jensen-Shannon divergence [11].

3 Datasets

CCTA Dataset. The CCTA dataset consists of 201 CCTA volumes gathered from eight different medical centers. Volumes were carefully selected and contain high quality scans with a low number of artefacts. It consists of 87 volumes of female and 114 volumes of male patients who are in average 69.3 $\pm$ 10.8 years old. Images are acquired from seven kinds of scanners: Siemens SOMATOM Force, Siemens SOMATOM Definition Flash, Siemens Sensation Cardiac 64, Canon Medical Systems Aquilion ONE, GE Medical Systems Discovery CT750 HD, GE Medical Systems Revolution Apex and GE Medical Systems Optima CT660. Volumes have a size of 512 $\times$ 512 $\times$ (130–420), with spacing 0.4–0.6 mm and a typical pixel size of around 0.33^2 mm^2. Landmarks include the position of the left and right coronary ostium and were manually annotated by trained radiologists. See Fig. 2 for example slices and annotations from the dataset. More examples are in the Appendix B.

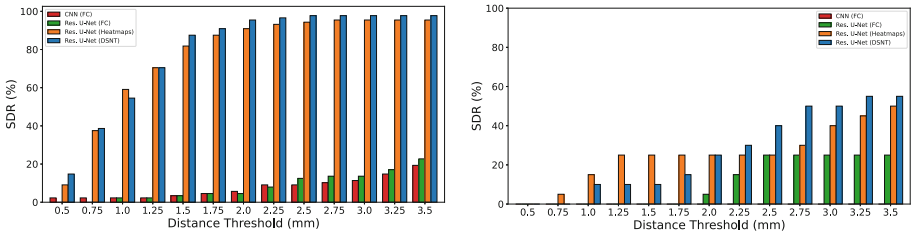
Table 1. Comparison of the Euclidean distance error (mm) between proposed method and other approaches on test volumes. Best results are bolded.

Model	Landmark	CCTA dataset				ImageTBAD dataset			
		Median	Mean	Std	IQR	Median	Mean	Std	IQR
CNN (FC)	Left CA ostium	6.70	6.84	2.98	3.82	8.38	9.43	4.20	4.98
	Right CA ostium	6.83	7.22	3.51	4.97	13.65	13.06	5.40	5.66
Residual U-Net (FC)	Left CA ostium	5.25	5.56	2.48	3.01	7.42	8.00	4.17	5.40
	Right CA ostium	5.60	6.08	2.99	4.43	11.91	12.12	6.10	8.45
Residual U-Net (Heatmaps)	Left CA ostium	1.21	1.45	1.23	0.76	3.63	3.85	1.44	2.00
	Right CA ostium	1.03	1.38	1.33	0.91	3.65	3.74	2.25	3.48
Residual U-Net (DSNT)	Left CA ostium	**1.14**	**1.18**	0.56	0.74	**3.48**	**3.49**	1.42	1.66
	Right CA ostium	**0.98**	**1.29**	1.16	0.82	**2.97**	**3.54**	1.88	2.71

CTA ImageTBAD Dataset. The ImageTBAD [29] is a public CTA image dataset of Type-B aortic dissection. Since the aortic non-ECG-gated CTA is not dedicated for coronary ostia imaging, the clarity of ostia localization in the CTA ImageTBAD dataset is significantly lower compared to the CCTA dataset. Despite that we decide to use it, as it is the only publicly available CT dataset known to the authors with good enough quality for the task of ostia localization. The dataset consists of 100 CTA images gathered from Guangdong Provincial Peoples' Hospital, China. It consists of 31 volumes of female and 69 volumes of male patients who are in average 52.5 ± 11.3 years old. Images are acquired from two kinds of scanners: Siemens SOMATOM Force, and Philips 256-slice Brilliance iCT system. Volumes have a size of $512 \times 512 \times (135–416)$. For information on spacing and voxel size please refer to [29]. The dataset consists of Type-B aortic dissection segmentation annotations, which are extended in this work to contain ostia landmarks. Identically to the CCTA dataset, these landmarks include the position of the left and right coronary ostium and are manually annotated by trained radiologists. Each annotation has been annotated upon agreement of at least two radiologists. During the annotation 23 volumes were discarded due to the insufficient data quality and reconstruction artefacts in the coronary ostium neighborhood resulting in 77 volumes with coronary ostia annotations. We share these annotations with the public for further research. See Fig. 2 for example slices and annotations from the dataset. More examples are in the Appendix *B*.

4 Experiments

We compare our method with three other regression methods on two datasets. Experiments are performed on each dataset separably. We do not mix the two datasets. For a fair comparison all methods consist of the same residual U-Net backbone. Details of the backbone architecture are in the Appendix *A*. Baseline methods are CNN (which consist of a U-Net encoding part) + fully connected (FC) layer for coordinates regression, Residual U-Net + FC for coordinates



a) CCTA dataset b) ImageTBAD dataset

Fig. 3. SDR for landmark localization. (a) CCTA dataset. (b) ImageTBAD dataset

regression, and a Residual U-Net Heatmap regression with argmax. We use a standard evaluation criteria [17] to compare models on the task of automatic landmark localization. These are the Euclidean distance (EUC) with its median, mean, standard deviation (STD), IQR and SDR. SDR is defined as the percentage of predicted landmarks which have the Euclidean distance to the reference landmark below the given threshold.

For each dataset all models are trained with a batch size of 4, a learning rate of 2.5×10^{-4} and Adam optimizer for $1,000$ epochs. Input volumes are resized to $144 \times 144 \times 144$ with trilinear interpolation. We augment the data during all trainings with rotations up to $\frac{\pi}{6}$ radians and translations up to 10 voxels in each dimension. To keep global volume context and resolve memory issues, volumes are downsized in the preprocessing step. Even though this process makes data less detailed and spatially accurate, we showcase that accuracy of results obtained from our proposed solution does not suffer and even surpasses other methods which incorporate local refinement. Since we are only interested in the heart area, we apply additional central cropping preprocessing step to each volume during training and inference for the ImageTBAD dataset. We implement methods with use of PyTorch and MONAI [2] libraries and use the NVIDIA GeForce GTX 1660 6GB for both training and inference. The average runtime and memory footprint analysis are in the Appendix *C*. For both datasets we use a cross-validation with 3-folds. The CCTA dataset is randomly divided into three folds of 104 training volumes, 53 validation volumes and one test set of 44 volumes. Because CTA ImageTBAD dataset has much fewer volumes, we use smaller validation and test sets. We randomly divide it into three folds of 57 training volumes, 10 validation volumes and one test set of 10 volumes.

Table 1 shows a comparison between our approach and other methods. The simplest methods, which use FC layer at the end of the backbone perform poorly. This proves that for the task of coronary ostia regression, or landmark regression in particular, utilizing heatmap generation is beneficial over exact coordinate regression. Using argmax over heatmaps or our proposed setup with DSNT leads to much higher results. Using the proposed method with DSNT, we achieve a median Euclidean distance error of 1.14 mm and 0.98 mm, a mean Euclidean

Table 2. Effect of the σ spread on the accuracy on the CCTA validation volumes.

Metric	Modeled heatmap standard deviation σ [voxel]								
	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Median [mm]	1.08	1.02	1.05	1.08	1.08	1.10	1.21	1.28	1.46
Mean [mm]	1.59	1.69	1.43	1.48	1.40	1.56	1.66	1.71	1.78
Std [mm]	2.35	3.00	1.79	1.79	1.29	1.80	1.64	1.83	1.45
IQR [mm]	0.89	0.89	0.86	0.85	0.78	0.93	1.05	1.09	1.20

distance error of 1.18 ± 0.56 mm and 1.29 ± 1.16 mm, and IQR of 0.74 mm and 0.82 mm on the left and right coronary ostium, respectively, on the 44 scans from the CCTA test set. On the ImageTBAD we achieve a median Euclidean distance error of 3.48 mm and 2.97 mm, a mean Euclidean distance error of 3.49 ± 1.42 mm and 2.97 ± 3.54 mm, and IQR of 1.66 mm and 2.71 mm on the left and right coronary ostium, respectively, on the 10 scans from the test set. On both datasets our proposed method achieves the highest or comparable result.

The SDR histogram plots are shown in Fig. 3. For the clinically accepted SDR range of 3.5mm, our method outperforms other approaches on the CCTA dataset achieving 97.73% SDR at 3.5 mm. The second best method achieves 95.45% SDR. On the ImageTBAD dataset, which consist of more challenging volumes, all methods perform much worse with our method achieving 55% SDR on the accepted range of 3.5 mm. The second best method achieves 50% SDR.

In an ablation study we evaluate the effect of σ in the Jensen-Shannon divergence regularization $\mathcal{L}_D$ which controls the desired heatmap spread over the volume on the CCTA dataset. See Table 2 for results. As expected, if σ is too large, or too small, heatmaps are forced to be shaped to unnatural sizes. This results in heatmaps having too much or too little context affecting the prediction of coronary ostium position. We find $\sigma = 3$ [voxel] to work best in our case.

5 Conclusion

In this study, we propose a one-step approach to the task of landmark localization from CT. We extend ImageTBAD CTA dataset annotations by coronary ostia landmarks and validate the method on them and on the CCTA dataset. Our solution shows promising results compared to methods, achieving mean Euclidean distance error of 1.18 ± 0.56 mm and 1.29 ± 1.16 mm on left and right coronary ostium respectively on the CCTA dataset and mean Euclidean distance error of 3.49 ± 1.42 mm and 3.54 ± 1.88 mm on left and right coronary ostium respectively on the ImageTBAD dataset. While on the CCTA dataset our proposed method achieves overall good results, there is a room for improvement for the ImageTBAD CTA dataset which is a more challenging dataset that consist of lower resolution images with a higher number of volumes with imaging artefacts.

References

1. Al, W.A., Jung, H.Y., Yun, I.D., Jang, Y., Park, H.B., Chang, H.J.: Automatic aortic valve landmark localization in coronary CT angiography using colonial walk. PloS one **13**(7), e0200317 (2018)
2. Consortium, M.: Monai: medical open network for AI (2020). https://doi.org/10.5281/zenodo.6114127, https://doi.org/10.5281/zenodo.6114127, If you use this software, please cite it using these metadata
3. Dabbah, M.A., et al.: Detection and location of 127 anatomical landmarks in diverse CT datasets. In: Medical Imaging 2014: Image Processing, vol. 9034, p. 903415. International Society for Optics and Photonics (2014)
4. Donner, R., Menze, B.H., Bischof, H., Langs, G.: Global localization of 3d anatomical structures by pre-filtered hough forests and discrete optimization. Med. Image Anal. **17**(8), 1304–1314 (2013)
5. Elattar, M., et al.: Automatic aortic root landmark detection in CTA images for preprocedural planning of transcatheter aortic valve implantation. Int. J. Cardiovascular Imaging **32**(3), 501–511 (2016)
6. Gao, Y., Shen, D.: Collaborative regression-based anatomical landmark detection. Phys. Med. Biol. **60**(24), 9377 (2015)
7. Han, D., Gao, Y., Wu, G., Yap, P.T., Shen, D.: Robust anatomical landmark detection for mr brain image registration. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8673, pp. 186–193. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10404-1_24
8. Ionasec, R.I., et al.: Dynamic model-driven quantitative and visual evaluation of the aortic valve from 4D CT. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008. LNCS, vol. 5241, pp. 686–694. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85988-8_82
9. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Adv. Neural Inform. Process. Syst. **28**, 2017–2025 (2015)
10. Khanna, A., Londhe, N.D., Gupta, S., Semwal, A.: A deep residual u-net convolutional neural network for automated lung segmentation in computed tomography images. Biocybern. Biomed. Eng. **40**(3), 1314–1327 (2020)
11. Lin, J.: Divergence measures based on the shannon entropy. IEEE Trans. Inform. Theory **37**(1), 145–151 (1991)
12. Lu, X., Jolly, M.-P.: Discriminative context modeling using auxiliary markers for lv landmark detection from a single MR image. In: Camara, O., Mansi, T., Pop, M., Rhode, K., Sermesant, M., Young, A. (eds.) STACOM 2012. LNCS, vol. 7746, pp. 105–114. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36961-2_13
13. Mahapatra, D.: landmark detection in cardiac MRI using learned local image statistics. In: Camara, O., Mansi, T., Pop, M., Rhode, K., Sermesant, M., Young, A. (eds.) STACOM 2012. LNCS, vol. 7746, pp. 115–124. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36961-2_14
14. Mostafa, A., Ghanem, A.M., El-Shatoury, M., Basha, T.: Improved centerline extraction in fully automated coronary ostium localization and centerline extraction framework using deep learning. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), pp. 3846–3849 (2021). https://doi.org/10.1109/EMBC46164.2021.9629655

15. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
16. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. arXiv preprint arXiv:1801.07372 (2018)
17. Noothout, J.M., et al.: Deep learning-based regression and classification for automatic landmark localization in medical images. IEEE Trans. Med. Imaging **39**(12), 4011–4022 (2020)
18. Noothout, J.M., de Vos, B.D., Wolterink, J.M., Leiner, T., Išgum, I.: CNN-based landmark detection in cardiac CTA scans. arXiv preprint arXiv:1804.04963 (2018)
19. Oktay, O., et al.: Stratified decision forests for accurate anatomical landmark localization in cardiac images. IEEE Trans. Med. Imaging **36**(1), 332–342 (2016)
20. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating spatial configuration into heatmap regression based CNNS for landmark localization. Med. Image Anal. **54**, 207–219 (2019)
21. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
22. Štern, D., Ebner, T., Urschler, M.: From local to global random regression forests: exploring anatomical landmark localization. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 221–229. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_26
23. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. Adv. Neural Inform. Process. Syst. **27**, 1799–1807 (2014)
24. Toshev, A., Szegedy, C.: Deeppose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660 (2014)
25. Urschler, M., Ebner, T., Štern, D.: Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. Med. Image Anal. **43**, 23–36 (2018)
26. Waechter, I., et al.: patient specific models for planning and guidance of minimally invasive aortic valve implantation. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6361, pp. 526–533. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15705-9_64
27. Wolterink, J.M., van Hamersvelt, R.W., Viergever, M.A., Leiner, T., Išgum, I.: Coronary artery centerline extraction in cardiac CT angiography using a CNN-based orientation classifier. Med. Image Anal. **51**, 46–60 (2019)
28. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1281–1290 (2017)
29. Yao, Z., et al.: Imagetbad: a 3d computed tomography angiography image dataset for automatic segmentation of type-b aortic dissection. Front. Physiol. **12** (2021)
30. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geosci. Remote Sens. Let. **15**(5), 749–753 (2018)
31. Zheng, Y., et al.: Automatic aorta segmentation and valve landmark detection in c-arm CT for transcatheter aortic valve implantation. IEEE Trans. Med. Imaging **31**(12), 2307–2321 (2012)



AMLP-Conv, a 3D Axial Long-range Interaction Multilayer Perceptron for CNNs

Savinien Bonheur[1,2(✉)], Michael Pienn[1], Horst Olschewski[1,3], Horst Bischof[2], and Martin Urschler[4]

[1] Ludwig Boltzmann Institute for Lung Vascular Research, Graz, Austria
[2] Institute of Computer Graphics and Vision, Graz University of Technology, Graz, Austria
bonheur@student.tugraz.at
[3] Department of Internal Medicine, Medical University of Graz, Graz, Austria
[4] School of Computer Science, University of Auckland, Auckland, New Zealand

Abstract. While Convolutional neural networks (CNN) have been the backbone of medical image analysis for years, their limited long-range interaction restrains their ability to encode long distance anatomical relationships. On the other hand, the current approach to capture long distance relationships, Transformers, is constrained by their quadratic scaling and their data inefficiency (arising from their lack of inductive biases). In this paper, we introduce the 3D Axial Multilayer Perceptron (AMLP), a long-range interaction module whose complexity scales linearly with spatial dimensions. This module is merged with CNNs to form the AMLP-Conv module, a long-range augmented convolution with strong inductive biases. Once combined with U-Net, our AMLP-Conv module leads to significant improvement, outperforming most transformer based U-Nets on the ACDC dataset, and reaching a new state-of-the-art result on the Multi-Modal Whole Heart Segmentation (MM-WHS) dataset with an almost 1.1% Dice score improvement over the previous scores on the Computed Tomography (CT) modality.

Keywords: MLP · Axial attention · Convolutional neural network · Multi-label · 3D semantic segmentation · Heart segmentation

1 Introduction

The encoding of long-range dependencies at the layer level is key to capture relationships between anatomical structures in many medical segmentation tasks and has been the focus of many publications in recent years. Earlier approaches tried to integrate long-range interaction as a module between convolutions [2,4,10,17,20,21] while recently, transformers [3,8,14,19,24] have been used to either complement convolutions or to replace them completely. Contrary to CNNs, transformer architectures are not biased toward local interaction [16]. Whereas this lack of bias allows transformers to consider images more holistically, this

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 328–337, 2022.
https://doi.org/10.1007/978-3-031-21014-3_34

comes at the cost of requiring more training data. While it is easier to find large annotated datasets for natural images, such as [13,18], the availability of large medical datasets is limited by both the variety of imaging modalities and the expertise needed to tackle the complexity of annotating these images. Thus, methods to augment CNNs with a non-local interaction while conserving their data efficiency are still needed.

Here, we introduce an alternative to transformers, Axial Multilayer Perceptron (AMLP) module, to capture long-range interaction and allow layer-wise context modulation in CNNs. Placed before each convolution to form the AMLP-Conv module, the AMLP modulates the input feature map by summarizing the information along each spatial axis. It considers the content and spatial information jointly to reweight each channel at each position independently, hence allowing axial context to influence information locally. Thus its complexity scales linearly with the input dimension. To prove the strength of long-range modeling, we propose the AMLP-Line, a pooling and downsampling free architecture (all layers are stacked sequentially at output resolution) which significantly outperforms U-Net on a CT MM-WHS [25] three-fold cross-validation experiment. Moreover, we show that combined with a U-Net architecture on a fairly big heart segmentation dataset, we are competitive with most transformer based approaches. Lastly, we show that the AMLP-Conv U-Net achieves a 1.1% improvement on the MM-WHS dataset challenge [25] state-of-the-art CT result.

2 Related Work

In the following, we review two types of approaches based on long-range dependencies. The first type of architecture, i.e. augmented convolution, uses feed forward layers and pre-convolution reweighting to dynamically modulate their input either channel-wise, pixel-wise, or both. The second type of architecture is based on transformers, where the input is split into non-overlapping patches, which are then compounded through a key-query interaction. The transformer architectures are often mixed with convolutions to benefit from the data efficiency arising from its local bias.

2.1 Augmented Convolution

Following the success of Squeeze and Excite (SE-Net) [10], where global pooling and MLP were combined in a channel attention module to globally reweight each feature channel, many fully connected layer and MLP-based attention mechanisms were proposed. For example in GCNet [2], SE-Net global average pooling is replaced by an adaptive pooling layer, while scSE [17] proposes to concurrently calculate spatial reweighting and channel reweighting. ECANet [20] introduces a one-dimensional convolutional layer as a more efficient alternative to fully connected layers. In CBAM [21], max pooling is added to complement SE-Net average pooling and a novel attention module, i.e. spatial attention, is introduced to leverage local context and create a scalar weighting map. Lastly, Chen et al. [4] propose to use a double attention, first generating a global context descriptor,

and then selecting a subset of global features adapted to each location content. All of the above methods use either context modulation alone [4,10,20,21], sequentially with spatial modulation [2], or concurrently as in [17]. However, joint consideration of spatial and channel modulation has not yet been applied.

2.2 Transformers

Following the recent success of transformers in vision [6], several transformer based approaches were proposed for semantic segmentation of medical images. Most approaches propose to mix transformers, utilizing their long-range interaction, and convolution, exploiting their strong local inductive biases. Inspired by [6], LeViT-UNet proposes to build an U-Net architecture with a transformer encoder and a convolutional decoder. Following on merging transformers and convolutions, TransUNet [3] propose to use transformer blocks to encode features at the bottom of an U-Net architecture. In [8], a transformer is used to encode the image before any convolution is used. And in [19], a local and a global branch are combined with a complex axial attention module. In nnFormer [24], transformer blocks are used in place of all convolutions with the exception of up- and downsampling layers. Finally, hierarchical transformers are proposed in SwinU-Net [14], to replace all convolutions with transformers.

While the transformers lack of visual biases is usually circumvented through pre-training on 2D natural images, it is unclear which biases and limitations this could introduce to 2D and 3D medical segmentation.

3 Method

Jointly applying spatial-wise and channel-wise modulation, our proposed AMLP module dynamically and uniquely reweights each channel at each location by summarizing the information across the whole image along each axis (i.e. height, width, depth). By axially acquiring information rather than globally, our AMLP can consider long range interaction while benefiting from a strong inductive bias (i.e. axial information is all you need) and limited complexity. Thus, its complexity scales linearly with the input size and the number of spatial dimensions.

$$AMLP(\mathbf{I_{d,h,w}}) = \sigma\left(\sum_{d=0}^{D-1} MLP_0(\mathbf{I_{d,h,w}}) + \sum_{h=0}^{H-1} MLP_1(\mathbf{I_{d,h,w}}) + \sum_{w=0}^{W-1} MLP_2(\mathbf{I_{d,h,w}})\right) \times \mathbf{I_{d,h,w}} \tag{1}$$

In Eq. 1, a 3D AMLP module applied on an input I of width W, height H, and depth D, applies three independent ReLU based MLPs to each voxel, before summing all values along the W axis, H axis, and D axis for the first, second and third MLP's output, respectively. The outputs of this operation are summed (broadcasting each MLP sum-collapsed dimension to match the input size) and passed to a sigmoid function σ to produce a position dependent (width-wise,

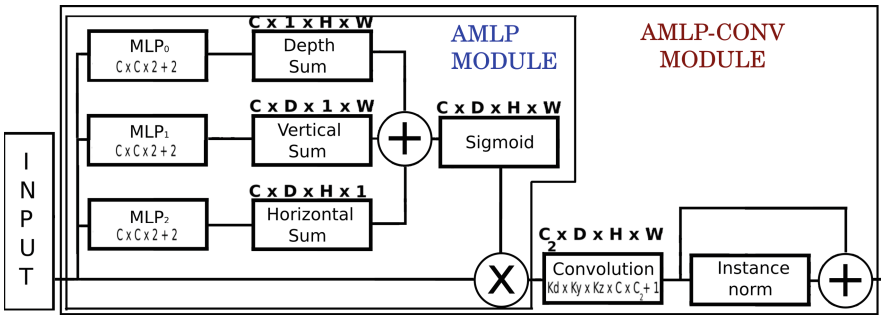


Fig. 1. A 3D AMLP-Conv module. Each block's output size is written above it, and each block's parameters count is written within it.

height-wise and depth-wise) and channel dependent weighting map, which is then multiplied with the input I.

To create an AMLP-Conv module (see Fig. 1) we simply pass the AMLP-Module to a convolution operation with window size Kd×Kh×Kw to form an output O. Lastly, an instance normalization operation is applied to O and added to a skip connection in order to stabilize training. The number of operations of our 3D AMLP-Conv module is $\mathrm{W} \times \mathrm{H} \times \mathrm{D} \times (\mathrm{P}_{conv} + 3\mathrm{P}_{mlp} + 6\mathrm{C}_{in} + \mathrm{P}_{\sigma} + \mathrm{P}_{normskip})$, where P_{σ} is the cost of the activation function, P_{conv} is the cost of a padded 3D convolution $(\mathrm{C}_{in} \times \mathrm{C}_{out} \times \mathrm{Kd} \times \mathrm{Kh} \times \mathrm{Kw} + 1 + \mathrm{P}_{\sigma})$, P_{mlp} the cost of an MLP $(\mathrm{C}_{in} \times \mathrm{C}_{out} \times 2 + 2 + \mathrm{P}_{\sigma})$, and $\mathrm{P}_{normskip}$ is the norm and skip connection's cost.

3.1 Datasets

ACDC [1]: This MRI dataset contains 100 patients (70 for training, 10 for validation, and 20 for testing). In each volume, the Myocardium (MV), the Left Ventricle (LV), and the Right Ventricle (RV) have been manually annotated by a single expert. All preprocessing is done using the pipeline from [24] for a final input size of $160 \times 160 \times 14$ (i.e. similar to nnFormer but smaller than nn-U-Net). The U-Net and AMLP-U-Net networks were built without batch normalization, pre-training or drop-out to isolate the AMLP-module's effect.

MM-WHS [25]: Introduced in MICCAI 2019 as a CT and MRI heart segmentation challenge, it contains 60 scans per modality, split into 20 training volumes and 40 testing volumes each. The segmentation masks contain the labels of 7 heart substructures, i.e. left ventricle, myocardium of the left ventricle, right ventricle, left atrium, right atrium, aorta and pulmonary artery. Acquired in several clinics and with different scanners, the scans come with different image quality, resolution, and voxel spacing. Using augmentation and preprocessing as presented in [15], each scan was resampled to a resolution of $64 \times 64 \times 64$ in the three-fold cross-validation experiment and to $96 \times 96 \times 96$ for the challenge itself. We used a spacing of $3 \times 3 \times 3$ for the CT scans and $4 \times 4 \times 4$ for the MRI scans.

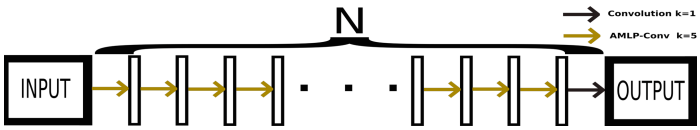


Fig. 2. AMLP-Line architecture

3.2 Evaluated Networks

ACDC: On this dataset we used the training strategy and augmentation from nnFormer pipeline[1] [24].

U-Net: this architecture, presents 64 feature maps in the first U-Net level (doubling at each sublevel).

AMLP-U-Net: this U-Net based architecture has 48 feature maps in the first level to match our U-Net parameter counts. Each 3D convolution module is replaced with our 3D AMLP-Conv module. While the spatial resolution of ACDC is anisotropic, we did not compensate for it while summing along axes.

MM-WHS: Averaged over 4 runs, all networks[2] has been trained on an Nvidia GeForce RT2080 Ti using the default AdaDelta optimizers [23] and the pre-processing pipeline from [15]. We calculate and report statistical significance with a p-value of 0.05.

U-Net: This is a classic U-Net architecture with 36 filters per feature map (64 for the challenge) and max pooling for downsampling. All convolutions have size $3 \times 3 \times 3$. Although U-Net usually has more parameters than the 454K present, the goal of this experiment is to propose a comparison with our AMLP-Conv U-Net with an equivalent parameters capacity. This network performs 100 iterations each ≈19.2 s and uses 1.6 GB of memory.

AMLP-Line: This architecture (see Fig. 2) uses 10 consecutive 3D AMLP-Conv modules with 54 filters and $5 \times 5 \times 5$ 3D convolution. Its smaller number of parameters (350K against 450K for the other network) is due to time constraints arising from the computation of each feature map at output resolution. This network performs 100 iterations each ≈85.5 s and uses 8.7 GB of memory.

AMLP-U-Net: The convolution modules of U-Net were replaced by 3D AMLP-Conv using $3 \times 3 \times 3$ convolutions. We use 27 filters per layer for the cross-validation experiment, and 48 filters during the challenge. The cross-validation network performs 100 iterations each ≈19.5 s and uses 2.6 GB of memory.

3.3 Results

Although we do not directly compare our method to augmented convolution methods, transformer approaches have been shown to outperform augmented convolution, either indirectly through the ImageNet benchmark [18] or with the comparison of GCNet [2] and Swin Transformer [14] on the Coco dataset [13].

[1] https://github.com/282857341/nnFormer.

[2] The code is available at https://github.com/savinienb/AMLP-Conv.

Table 1. ACDC dataset. Multi-label Dice scores in % on the ACDC dataset.

Network	RV	LV	Myo	Mean
TransUNet [3]	88.86	84.54	95.73	89.71
SwinU-Net [14]	88.55	85.62	95.83	90.00
LeVit-UNet0384 [22]	89.55	87.64	93.76	90.32
nnU-Net(3D) [11]	**90.25**	89.10	95.41	91.59
nnFormer [24]	90.22	**89.53**	**95.59**	**91.78**
U-Net	88.79	88.89	95.07	90.91
AMLP-U-Net	89.10	88.73	95.10	90.98

Table 1 shows that in the ACDC dataset, the AMLP-U-Net is competitive with most transformer architectures despite their use of pre-training.

While the AMLP-U-Net does not reach the performances of nnFormer, the performance difference between their nn-U-Net and our U-Net seems to indicate that optimizations could close the gap.

Table 2 shows that on CT images, both our AMLP-Line (a down- and upsampling free architecture) and AMLP-U-Net significantly outperform U-Net. On the MRI images however, U-Net outperforms AMLP-U-Net.his may comefrom a less direct linkage of intensity values to material composition (comparedto CT) and a small training dataset (13 training scans per fold). This modalitygap, seems to reduce with more training data (i.e. 20 scans) as seen in Table 3.

Table 2. MM-WHS dataset - CT & MRI cross-validation. Multi-label Dice scores of the evaluated networks in % (mean ± standard deviation) using a three-fold cross-validation averaged over 4 runs on the training images of the MM-WHS challenge. The input images' modality is indicated the "Data Type" column. Results shown for the individual cardiac structures (LV - left ventricle, Myo - left ventricle myocardium, RV - right ventricle, LA -left atrium, RA - right atrium, aorta - ascending aorta, PA - pulmonary artery and the mean Dice score over the seven heart substructures). The number of network parameters is shown as multiples of thousands.* indicate statistical significance $p < 0.05$.

Data type	Network	#Params	LV	RV	LA	RA	Myo	aorta	PA	Mean
CT	U-Net	454K	93.01* (±3.07)	89.35 (±3.82)	92.12 (±4.00)	88.04 (±5.77)	88.80* (±3.39)	87.02 (±11.21)	84.10 (±10.12)	88.92* (±3.72)
	AMLP-Line	347K	93.44 (±3.09)	**90.41** (±3.18)	**92.57** (±3.71)	88.24 (±5.65)	89.81 (±3.97)	**95.17** (±2.11)	85.57 (±8.63)	**90.75** (±2.38)
	AMLP-U-Net	445K	**93.60** (±2.98)	89.88 (±3.33)	92.49 (±3.81)	**88.52** (±5.45)	**89.99** (±3.22)	94.49 (±4.32)	**86.15** (±7.55)	90.73 (±2.29)
MRI	U-Net	454K	91.12 (±4.02)	83.41 (±9.80)	**86.99** (±1.38)	87.69 (±4.76)	80.28 (±5.78)	**86.04** (±3.45)	**77.76*** (±1.83)	**84.76** (±3.44)
	AMLP-Line	347K	**91.96** (±4.28)	**84.90** (±9.55)	84.56 (±4.72)	**87.84** (±5.49)	80.52 (±3.45)	79.27 (±19.97)	74.91 (±4.27)	83.42 (±4.46)
	AMLP-U-Net	445K	91.20 (±3.25)	84.12 (±6.38)	84.27 (±2.50)	86.95 (±3.29)	**81.94** (±4.17)	80.31 (±24.11)	76.60 (±4.75)	83.63 (±3.91)

Table 3. MM-WHS challenge. Multi-label Dice scores (Dice), Jaccard score (Jaccard), Surface Distance (SD) and Hausdorff distance (HD) on the MM-WHS challenge set for both CT and MRI scans.* indicate statistical significance $p<0.05$ between our two networks.

	Teams	Dice	Jaccard	SD (mm)	HD (mm)
CT	[7]	90.4 ± 5.2	0.82 ± 0.06	1.11 ± 0.02	**15.36 ± 12.37**
	[12]	91.4 ± 0.08	0.84 ± 0.08	1.42 ± 0.46	28.04 ± 12.14
	[15]	90.8 ± 8.6	0.83 ± 0.04	1.12 ± 0.25	25.24 ± 10.81
	U-Net (ours)	91.8 ± 2.2	0.85 ± 0.03	0.91 ± 0.18	17.18 ± 9.63
	AMLP-U-Net (ours)	**92.5* ± 1.6**	**0.86* ± 0.03**	**0.89 ± 0.20**	18.10 ± 11.75
MRI	[5]	89.2 ± 3.6	**0.81 ± 0.04**	1.8 ± 0.9	24.21 ± 18.89
	[7]	**89.5 ± 2.2**	**0.81 ± 0.08**	1.86 ± 0.67	25.66 ± 16.40
	[12]	89.0 ± 7.5	0.80 ± 0.08	1.42 ± 0.46	28.04 ± 12.14
	[15]	86.3 ± 4.3	0.76 ± 0.06	1.89 ± 0.78	30.23 ± 14.05
	U-Net (ours)	86.8 ± 3.5	0.77 ± 0.05	1.42 ± 0.46	24.69 ± 17.22
	AMLP-U-Net (ours)	87.0 ± 3.7	0.78 ± 0.05	1.48 ± 0.86	25.34 ± 18.72

In Table 3, our AMLP network slightly improves the U-Net Dice score results on the MRI modality and outperforms all other approaches for CT images with a Dice score improvement of ∼1.1% over other teams' results (with a significant ∼0.7% Dice score increase over our U-Net).

Figure 3 shows the segmentations with the best and worst Dice scores for the CT and MRI scans, respectively. There is an over-segmentation of the right atrium (in red) by both the AMLP-U-Net and the AMLP-Line network in the worst CT modality result (second row). On the MRI images with the result (fourth row), while both our U-Net and our AMLP-U-Net wrongly labeled the ascending Aorta (in violet) as the left ventricle (dark blue), the AMLP-Line network does not make this mistake. Moreover, the AMLP module reduces the segmentation error for the right ventricle (in green), even picking up its second part (see red arrow).

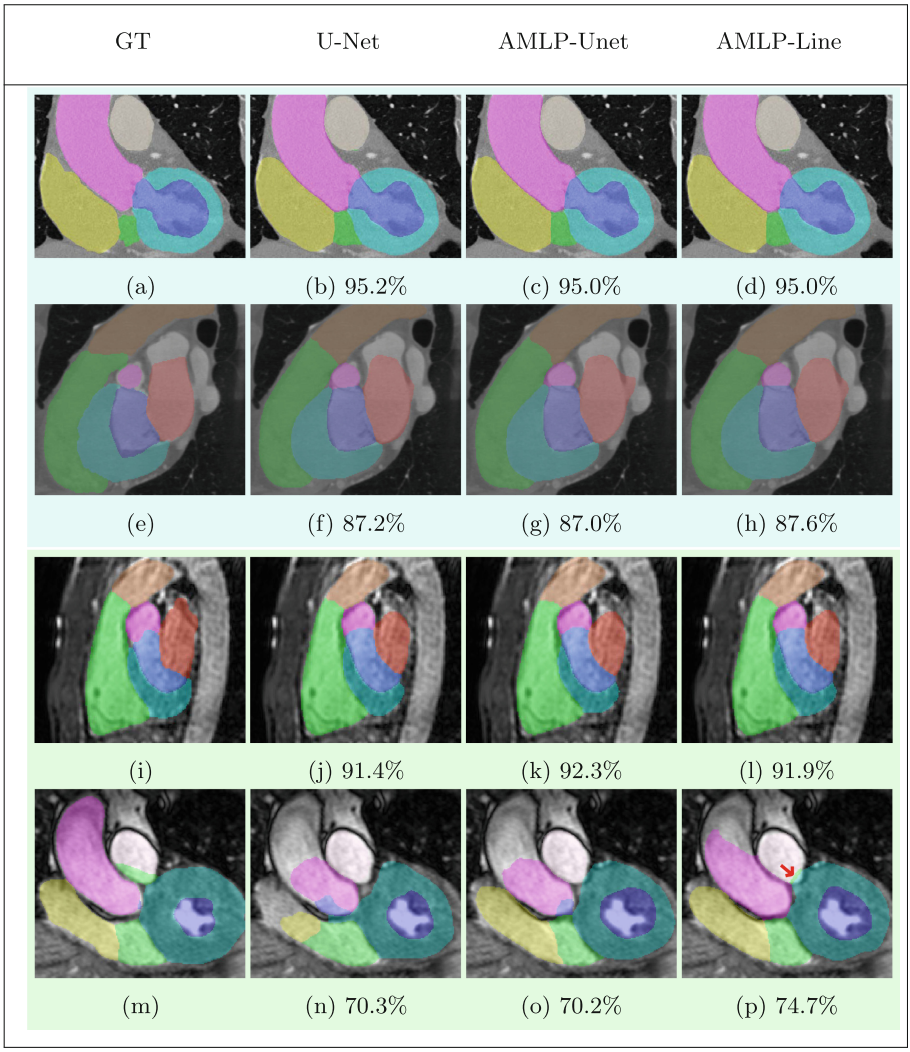


Fig. 3. Best (row 1) and worst (row 2) CT (in cyan) and best (row 3) and worst (row 4) MRI (in green) segmentation and Dice score for U-Net and our AMLP-U-Net. (Color figure online)

4 Conclusion and Future Works

In this paper, we introduced the AMLP-Conv module for semantic segmentation. Scaling linearly in complexity with the input size, our AMLP module encodes long-distance relationships through a combined channel and spatial integration and shows itself to be a powerful complement to convolution. Indeed, our AMLP-Line model, a down- and upsampling-free model, significantly outperform

U-Net on the CT MM-WHS three-fold cross-validation experiment. Moreover, our AMLP-U-Net is competitive with most transformer baselines on the ACDC dataset, and increases by +1.1% Dice score the MM-WHS challenge current state-of-the-art CT results. While the benefit of our method seems limited for MRI imaging, we believe it to stem from the high variability in MRI's voxel intensity. We, therefore, believe that research should investigate a regularization scheme to allow AMLP-Conv module CT's improvement to translate to the MRI modality. As well, further research should focus on visualizing the feature maps for both AMLP-UNet and AMLP-Line to evaluate how different architecture biases influence the network's internal representation.

References

1. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A. et al.: deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. In: IEEE Transactions on Medical Imaging, vol. 37 (2018)
2. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: GCNet: non-local Networks meet Squeeze-Excitation Networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
3. Chen, J., et al.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J.: A2-Nets: double attention networks. Adv. Neural Inform. Process. Syst. **31** (2018)
5. Dong, S., et al.: DeU-Net 2.0: enhanced deformable U-Net for 3d cardiac cine MRI segmentation. Med. Image Anal. **78** (2022)
6. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
7. Habijan, M., Galić, I., Leventić, H., Romić, K.: Whole heart segmentation using 3D FM-pre-resnet encoder-decoder based architecture with variational autoencoder regularization. Appl. Sci. **11** (2021)
8. Hatamizadeh, A., et al.: UNETR: transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2022)
9. Heinrich, M.P.: Intra-operative ultrasound to MRI fusion with a public multimodal discrete registration tool. In: Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation (2018)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
11. Isensee, F., et al.: NNU-Net: self-adapting framework for U-Net-based medical image segmentation. Nat. Methods **18** (2021)
12. Liao, X., Qian, Y., Chen, Y., Xiong, X., Wang, Q., Heng, P.A.: MMTLNet: multi-modality transfer learning network with adversarial training for 3D whole heart segmentation. Comput. Med. Imaging Graph. **85** (2020)
13. Lin, T.Y., et al.: Microsoft coco: common objects in context. In: European Conference on Computer Vision (2014)

14. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
15. Payer, C., Štern, D., Bischof, H., Urschler, M.: Multi-label whole heart segmentation using anatomical label configurations. In: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. 2017, vol. 10663 (2018)
16. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Adv. Neural Inform. Process. Syst. **34** (2021)
17. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2018)
18. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115** (2015)
19. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: medical transformer: gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2021)
20. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: efficient channel attention for deep convolutional neural networks, 2020. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2020)
21. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
22. Xu, G., Wu, X., Zhang, X., He, X.: LeViT-UNet: Make Faster Encoders with Transformer for Medical Image Segmentation. arXiv preprint arXiv:2107.08623 (2021)
23. Zeiler, M.D.: ADADELTA: An Adaptive Learning Rate Method. arXiv preprint arXiv:1212.5701 (2012)
24. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnFormer: Interleaved Transformer for Volumetric Segmentation. arXiv preprint arXiv:2109.03201 (2021)
25. Zhuang, X., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Med. Image Anal. **58** (2019)



Neural State-Space Modeling with Latent Causal-Effect Disentanglement

Maryam Toloubidokhti(✉), Ryan Missel, Xiajun Jiang, Niels Otani, and Linwei Wang

Rochester Institute of Technology, Rochester 14623, USA
{mt6129,rxm7244,xj7056,nfosma,lxwast}@rit.edu

Abstract. Despite substantial progress in deep learning approaches to time-series reconstruction, no existing methods are designed to uncover local activities with minute signal strength due to their negligible contribution to the optimization loss. Such local activities however can signify important abnormal events in physiological systems, such as an extra foci triggering an abnormal propagation of electrical waves in the heart. We discuss a novel technique for reconstructing such local activity that, while small in signal strength, is the cause of subsequent global activities that have larger signal strength. Our central innovation is to approach this by explicitly modeling and disentangling how the latent state of a system is influenced by potential *hidden internal interventions*. In a novel neural formulation of state-space models (SSMs), we first introduce causal-effect modeling of the latent dynamics via a system of interacting neural ODEs that separately describes 1) the continuous-time dynamics of the internal intervention, and 2) its effect on the trajectory of the system's native state. Because the intervention can not be directly observed but have to be disentangled from the observed subsequent effect, we integrate knowledge of the native *intervention-free* dynamics of a system, and infer the hidden intervention by assuming it to be responsible for differences observed between the actual and hypothetical *intervention-free* dynamics. We demonstrated a proof-of-concept of the presented framework on reconstructing ectopic foci disrupting the course of normal cardiac electrical propagation from remote observations.

Keywords: Neural ODE · Intervention modelling · Cardiac EP

1 Introduction

Advancements in modern deep learning have resulted in substantial progress in our ability to model, reconstruct, and forecast complex time-series data, most notably through recurrent neural networks (RNNs) and neural ordinary differential equations (neural ODEs) [1,13,15,16]. Many of these techniques have been

M. Toloubidokhti and R. Missel—Contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_35.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 338–347, 2022.
https://doi.org/10.1007/978-3-031-21014-3_35

successfully adapted to model the spatiotemporal dynamics of complex physiological systems, such as in reconstructing [5,9] or forecasting the electrical activity of the cardiac system [11]. While strong in modeling general system dynamics, however, these methods are not designed to uncover local activity with minute signal strength compared to the global dynamics. Such local activities are common and often signify important abnormal events in physiological systems, such as an extra foci that triggers an abnormal propagation of electrical waves amid normal propagation patterns (Fig. 3). Due to negligible optimization loss contribution given weak signal strength, reconstructing these local activities remains challenging despite deep learning advances in time-series modelling.

We present a fundamentally novel perspective of this problem: instead of general reconstruction, we stress that these local activities, while small in signal strength, are the cause of the subsequent global activities that have larger signal strength. Therefore, we may be able to disentangle the *cause* and *effect* of the latent dynamics by explicitly modeling and inferring how the latent state of a system is influenced by potential hidden internal *interventions*. Modeling the effect on system dynamics by exogenous inputs is well-studied in classic state-space systems (SSMs), and has seen recent successes in neural-network modeling of SSMs [3,6,10,12]. These works, however, focus on external interventions that are either known or directly observed, whereas we consider hidden and internal causes of system's abnormal events. As these internal causes are not directly observable, we propose to disentangle them from the observed effects on the system's dynamics. Rather than learning the complex intervened dynamics of a system from scratch as in existing works [6], we incorporate knowledge of the native *intervention-free* dynamics of a system and focus only on learning how it is influenced by potential hidden interventions.

To this end we present a novel neural SSM with two major innovations. First, drawing inspirations from related work in intervention modeling [6], we introduce causal-effect modeling of the latent dynamics via a system of interacting neural ODEs that separately describe 1) the continuous-time dynamics of the internal intervention, and 2) its effect on the trajectory of the system's native state. Second, to disentangle the cause and effect from their collective observations, we 1) leverage data of intervention-free systems to pre-train a neural ODE of *native* dynamics and integrate it into the intervention ODEs, and 2) infer the hidden intervention by assuming it to be responsible for differences observed between the actual and hypothetical *intervention-free* dynamics at each time frame.

We demonstrated a proof-of-concept of the presented framework on reconstructing ectopic foci that disrupt normal electrical propagation using remote observations. We compared the presented method to the two most relevant time-series modeling approaches: 1) a global neural ODE to describe latent system dynamics, and 2) a neural SSM with a latent neural ODE lacking intervention modeling, developed for reconstructing cardiac electrical propagation [9]. Experiments were conducted on synthetic data with controlled *internal interventions*, simulated in 2D settings. Our results indicated that the presented method delivers more accurate inverse estimations in terms of localizing the triggering events.

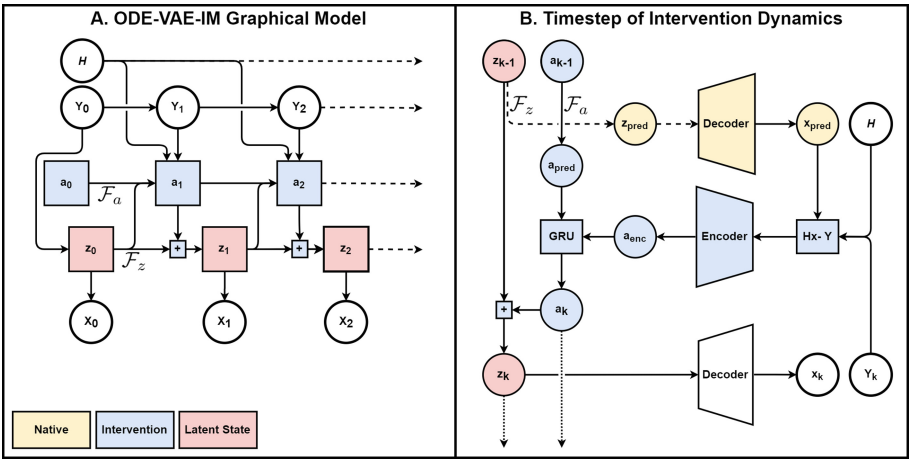


Fig. 1. A) Graphical overview of the proposed network. Inputs $\mathbf{H}$ and $\mathbf{Y}_{0:T}$ are given to the intervention dynamics $\mathcal{F}_a$ and influence the propagation of $\mathcal{F}_{(z,a)}$. B) Schematic of a single intervention step, showing the estimation of the intervention latent variable $\mathbf{a}_k$ from data-space observations and *hypothetical intervention-free observations*.

2 Methodology

We formulate a neural SSM where the system dynamics is described on a lower-dimensional latent manifold in separation from its emission to the data space as illustrated in Fig. 1. Consider cardiac electrical propagation $\mathbf{x_t}$ and its body-surface measurement $\mathbf{Y_t}$ with a physics-based relation $\mathbf{Y_t} = \mathbf{H}\mathbf{x}_t$. We enable causal-effect intervention modelling by sequentially learning two separate neural ODE functions, $\mathcal{F}_z$ and $\mathcal{F}_a$, that describe *native* system dynamics and causal dynamics, respectively. The effect of intervention $\mathbf{a}$ on $\mathbf{z}$ is modelled through a coupled neural ODE function, $\mathcal{F}_{z,a}$. Because intervention $\mathbf{a}$ is not directly observable, it is not possible to separately learn $\mathcal{F}_z$, $\mathcal{F}_a$ and $\mathcal{F}_{z,a}$ all from scratch from only their collective observations. Instead, we poise that we must leverage knowledge of native dynamics $\mathcal{F}_z$, in order to be able to generate *hypothetical observation* $\hat{\mathbf{Y}}_t = \mathbf{H}\hat{\mathbf{x}}_t$ assuming intervention-free dynamics since last observation $\mathbf{Y}_{t-1}$. Then using the residual between the actual observation $\mathbf{Y}_t$ and hypothetical observation $\hat{\mathbf{Y}}_t$ allows us to model and estimate intervention $\mathbf{a}_t$.

A two-stage optimization process is leveraged: First, the *native* dynamics of the system are learned on an *intervention-free* subset of data (termed *ODE-VAE*) which is then statically integrated into a system of intervention ODEs optimized on the intervention set (termed *ODE-VAE-IM*).

In Sects. 2.1–2.2 we describe the native dynamics model, the predictive model enabling causal-effect learning, the estimation of the latent intervention variable through global observations, and the model optimization loop.

2.1 Modeling and Learning Intervention-Free Native Dynamics

An ODE function $\mathcal{F}_z$ is designed to learn the *native* dynamics function that handles the propagation of the system in the absence of interventions. Since we leverage this function to hypothesize the intervention-free native state of a system given at any previous system state, it must have a strong long-term forecasting ability. Therefore, inspired by [18], 1.) an encoding network Enc_z initializes the latent ODEs vector field $\mathbf{z}_0$ using the first few frames of sequential input $\mathbf{Y}_{0:k}$ (Eq. 1), 2.) an initial value problem is solved up to time $\mathbf{T}$ using the ODE function $\mathcal{F}_z$ (Eq. 2), and 3.) a decoding network Dec_z converts the latent trajectory $\mathbf{z}_{0:T}$ to the output space $\mathbf{X}_{0:T}$ (Eq. 3):

$$\mathbf{z}_0 = Enc_z(\mathbf{Y}_{0:k}) \tag{1}$$

$$\mathbf{z}_i = \mathbf{z}_{i-1} + \int_{t_{i-1}}^{t_i} \mathcal{F}_z(\mathbf{z}_\tau) d\tau \tag{2}$$

$$\hat{\mathbf{x}}_i = Dec_z(\mathbf{z}_i), \tag{3}$$

where Enc_z and Dec_z are represented by CNNs and the sequence input length k is empirically tuned.

We train this native dynamics model on datasets of *intervention-free* systems. After initializing the latent state, the full trajectory is predicted using $\mathcal{F}_z$, from which the reconstructed $\hat{\mathbf{x}}$ is compared with the ground truth in the loss:

$$\mathcal{L}_{native}(\mathbf{x}_{0:T}, \hat{\mathbf{x}}_{0:T}) = \beta[(\mathbf{x}_0 * log(\hat{\mathbf{x}}_0)) + (1 - \hat{\mathbf{x}}_0) * log(1 - \hat{\mathbf{x}}_0)] \tag{4}$$

$$+ \frac{1}{T} \sum_{i=1}^{T} (\mathbf{x}_i * log(\hat{\mathbf{x}}_i) + (1 - \hat{\mathbf{x}_i}) * log(1 - \hat{\mathbf{x}}_i), \tag{5}$$

where binary cross entropy (BCE) is applied to each frame and β represents a loss weighting coefficient. We explicitly separate out the reconstruction of $\mathbf{x}_0$ to emphasize the accuracy of the initial conditions, $\mathbf{z}_0$, independent of the accuracy of the dynamic transition models. Because the network is asked to reconstruct the full sequence of $\mathbf{x}_{0:T}$ from only an estimated initial condition $\mathbf{z}_0$, it promotes learning strong, continuous dynamics models $\mathcal{F}_z$ that can simulate the system given any initial condition of the space. This function, once trained, will be incorporated into the intervention-effects function $\mathcal{F}_{(z,a)}$ as a static dynamics function whose trajectory is influenced by the intervention state.

2.2 Modelling and Learning Intervention Dynamics

The system dynamics predictive model is described by two ODE functions, $\mathcal{F}_a$ and $\mathcal{F}_{(z,a)}$, that aim to separately estimate the cause and effect of interventions in the latent space, respectively. To achieve this, we use the state-space Bayesian filtering setting where the system propagation undergoes a series of predict-from-dynamics and update-from-observations steps [3,4,10]. The former comprises of the predictive model which, given a current state $\mathbf{s}_{t-1}$, predicts the next temporal state, $\hat{\mathbf{s}}_t$, using the transition function $\mathcal{F}$. The latter corrects the temporal

prediction through an update function, $\mathcal{G}$, using an inferred state $\mathbf{s}_{enc}(t)$ from the current observation $\mathbf{Y}_i$.

Predictive System Dynamics Model. $\mathcal{F}_a$ handles the prediction of the intervention state, a_t, while $\mathcal{F}_{(z,a)}$ handles the intervention-influenced prediction of the latent state z_t, denoted by the equations in Eq. 6, respectively.

$$\textbf{Prediction: } \begin{bmatrix} \hat{\mathbf{a}}_i \\ \hat{\mathbf{z}}_i \end{bmatrix} = \begin{bmatrix} \mathbf{a}_{i-1} \\ \mathbf{z}_{i-1} \end{bmatrix} + \int_{t_{i-1}}^{t_i} \begin{bmatrix} \mathcal{F}_a(\mathbf{a}_\tau) \\ \mathcal{F}_z(\mathbf{z}_\tau) + \mathcal{F}_a(\mathbf{a}_\tau) \end{bmatrix} d\tau \tag{6}$$

Estimating the Intervention Latent Variable. Different from Sect. 2.1 wherein our goal is to learn a native dynamic model that has strong standalone forecasting ability, here our interest is to extract information from available observations at each time frame to uncover hidden interventions. A dynamical function describing the temporal cause of intervention alone is not enough to produce a meaningful causal-effect model. Rather, an observational source from which the latest intervention information can be inferred is required [6,19]. We hypothesize that the latent variable associated with the interventions rises from differences observed between data space observations and *hypothetical observations* continuing under *intervention-free* dynamics at each timestep. For inverse image reconstruction problems, this can be formulated as the element-wise difference in the inverse solution difference of $H\hat{\mathbf{x}}_t^{\mathbf{z}}$ and $\mathbf{Y}_t$. This difference is then used as input to an encoder to output the estimated latent intervention state.

$$\textbf{Update: } \mathbf{a}_{enc} = Enc_a(||\mathbf{H}\hat{\mathbf{x}}_i - \mathbf{Y}_i||_{i:(i+k)}) \tag{7}$$

$$\mathbf{a}_i = \mathcal{G}_a(\hat{\mathbf{a}}_i, \mathbf{a}_{enc}) \tag{8}$$

We make use of the Gated Recurrent Unit (GRU) cell [2], $\mathcal{G}_a$ to learn a weighted combination of the encoded intervention variable and the prediction coming from the intervention dynamics function. A number of methods may be applied to improve the forward intervention estimation. One such choice, and the one utilized in this work, is to predict a number of timesteps ahead and get the inverse solution difference over a sequence of frames rather than just the current predicted frame. This provides increased temporal information to the intervention encoder by exposing the longer-term differences that would arise from the current state.

Optimizing *Intervention* Models. The loss function is the data reconstruction loss averaged over each time frame:

$$\mathcal{L}_{intv}(\mathbf{X}_{0:T}, \hat{\mathbf{X}}_{0:T}) = \frac{1}{N * T} \sum_{n=0}^{N} \sum_{i=0}^{T} (\mathbf{x}_{n,i} * log(\hat{\mathbf{x}}_{n,i}) + (1 - \hat{\mathbf{x}_{n,i}}) * log(1 - \hat{\mathbf{x}}_{n,i}), \tag{9}$$

where BCE is applied to each time frame. As we are no longer care about the solution of an initial value problem, the emphasis on $\mathbf{x}_0$ reconstruction is lifted.

3 Experiments

For the task of reconstructing ectopic foci that disrupt normal electrical propagation using remote observations, we include three baseline methods: *ECGI.*)

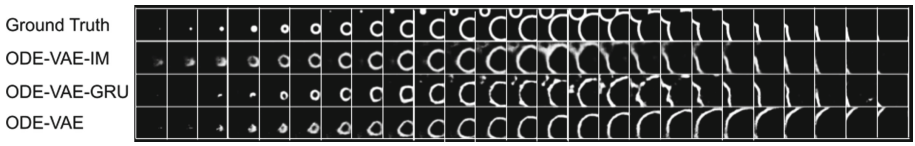


Fig. 2. Synthetic TMP per-step reconstruction in which additional excitations occur.

a first order Tikhonov regularization for reconstruction, *ODE-VAE.*) the direct application of the native dynamics network to the given intervention set, and *ODE-VAE-GRU.*) a Bayesian filtering ablation where a latent update mechanism (GRU) is applied at every timestep using the next 3 observations straight as input. This represents a model optimized on native and intervention dynamics together without clear disentanglement or ability to preserve *intervention-free* dynamics. We denote the proposed method as *ODE-VAE-IM* throughout the experiments. The observation data $\mathbf{X}_{0:T}$ is passed through a Bernoulli filter between activated and deactivated nodes following min-max normalization and all network decoders have a Sigmoid activation on the output images. The intervention variable encoder (Enc_a) has non-activated outputs (Fig. 2).

Initial learning rates were found using the cyclical learning rate estimation technique [17] and are decayed by 0.5 at set points throughout training. We used AdamW optimizer [14] with a weight decay of $1e{-}2$, a batch size of 16, and a latent dimension of 12 across all methods. All experiments were run on NVIDIA Tesla T4s with 16GB memory, taking $\sim$8 h to train. Our implementation and saved models are available at https://github.com/qu-gg/causal-effect-neural-ssm, along with more examples in Supplementary Material.

Data Generation. Following the Fitzhugh-Nagumo model [7], we simulated the transmembrane potentials on a 100 * 100 mm 2D grid. This synthetic dataset includes two subsets 1.) native transmembrane potential set containing 1000 voltage maps, where the initial excitation location is chosen randomly across the grid and 2.) transmembrane potentials in the presence of an extra focci. Simulating the setting where the extra Foci intervenes in the normal dynamics, we generated a dataset of 705 samples with varying initial excitation location and time and location of the extra Foci.

3.1 Dynamics Reconstruction

In this section, we showcase the reconstructive and temporal capabilities of the method on normal and intervention dynamics.

Normal Dynamics. We first highlight the *naive* dynamics model's performance on the non-intervention datasets used for pre-training. To ensure that a robust vector field is trained, our base dataset combines both single and double excitation propagations that occur within the first few frames. We posit that by introducing more propagation patterns in the base dynamics, even if they are not used in intervention, helps the ODE function and emission model to handle

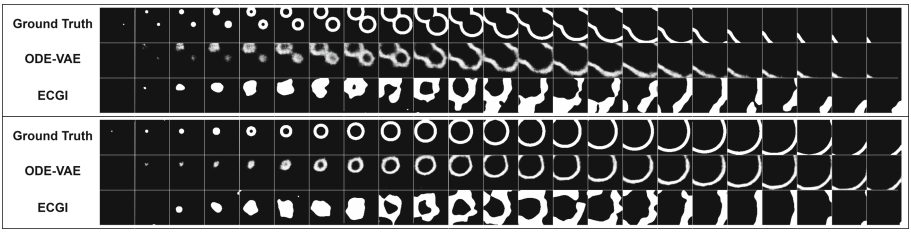


Fig. 3. Reconstruction of electrical propagation in which no interventions (foci) occur.

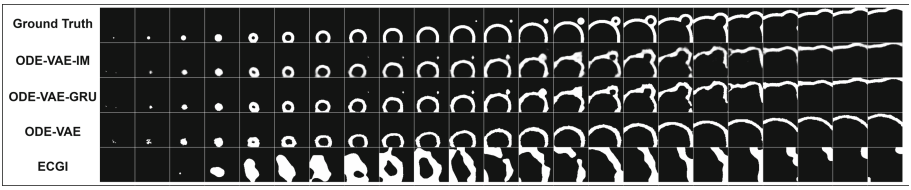


Fig. 4. Reconstruction of electrical propagation in which ectopic foci occurs.

Table 1. Comparison in identifying the intervention foci's location and activation timestep. Identification percentage represents cases with clear foci reconstruction.

Model	% Foci identified	Timestep activation MAE	Foci location error
ODE-VAE-IM	0.85	3.56	54.90
ODE-VAE-GRU	0.79	3.92	53.71
ODE-VAE	0.39	3.07	57.28
ECGI	0.52	4.10	52.35

unstable vector state jumps. Figure 3 highlights two samples of per-step reconstructions, showcasing that strong dynamics are established.

Ectopic Foci. Figure 4 highlights per-timestep reconstructions of a single sample across all methods. *ODE-VAE*, the base dynamics model with no means of intervention, manages to reconstruct a propagated waveform of the *native* dynamics. Leveraging this fixed native dynamics function, the presented *ODE-VAE-IM* is able to uncover the hidden intervention dynamics.

3.2 Localization Results

To gauge the proposed method's numeric performance, we performed a localization test in which the time and location of interventional foci were identified from the reconstruction results across a subset of 100 samples. The mean absolute error (MAE) in time frames, the Euclidean distance in location, and the percentage of identified foci are used for quantitative metrics. Table 1 showcases

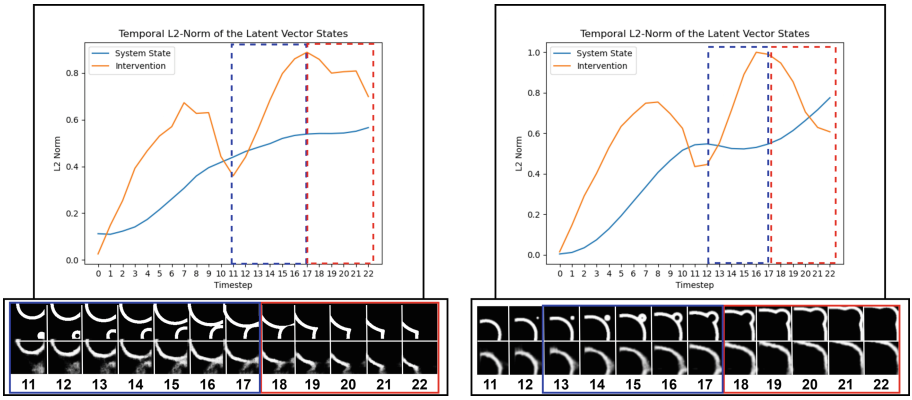


Fig. 5. Visualizations of the L2-Norm of system and intervention states over time.

these results, in which the ability of *ODE-VAE-IM* in identifying the presence of abnormal foci is highlighted.

3.3 Latent Norm Ablation

We performed an ablation study on the latent dynamics of the cause and effect by visualizing the L^2-Norm of their respective latent vector states over time for two examples in Fig. 5. Temporal windows are highlighted in which the intervention dynamics correctly capture the windows of starting and ending intervention activity. To facilitate comparisons between each latent space, we first perform min-max normalization over the L^2-Norms of each latent space to bring them within the same data range.

4 Conclusion

In this work, we propose a proof-of-concept interventional modelling framework to tackle the problem of low-strength local activity that signifies important abnormal triggering events in a dynamic system. We introduce a two-ODE system that separately models the cause and effect of system's latent state under the influence of hidden internal interventions. To this end we leverage pre-trained dynamical functions describing *intervention-free* native dynamics of a system. We demonstrated the frameworks performance on reconstruction of the ectopic foci causing abnormality in cardiac electrical activity.

Future Work. Future work aims to extend this framework into 3-D inverse reconstruction and pacing localization using graph neural networks to enable experimentation on clinical data, following [8,9]. One promising direction is learning multi-interventional dynamics simultaneously via attention mechanisms.

Limitations. The transition between training $\mathcal{F}_z$ and $\mathcal{F}_a$ has a period of re-training up to the original performance, which identity-based initialization or loss-influence annealing $\mathcal{F}_a$ may alleviate. Enc_z has potential generalization problems when the initial frame distributions shift between native and intervention sets and requires experimentation on its general training setup.

References

1. Chen, R.T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.: Neural ordinary differential equations (2019)
2. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation (2014)
3. De Brouwer, E., Simm, J., Arany, A., Moreau, Y.: Gru-ode-bayes: continuous modeling of sporadically-observed time series. Adv. Neural Inform. Process. Syst. **32** (2019)
4. Fraccaro, M., Kamronn, S., Paquet, U., Winther, O.: A disentangled recognition and nonlinear dynamics model for unsupervised learning. Adv. Neural Inform. Process. Syst. **30** (2017)
5. Ghimire, S., Dhamala, J., Gyawali, P.K., Sapp, J.L., Horacek, M., Wang, L.: Generative modeling and inverse imaging of cardiac transmembrane potential. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 508–516. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_57
6. Gwak, D., Sim, G., Poli, M., Massaroli, S., Choo, J., Choi, E.: Neural ordinary differential equations for intervention modeling. arXiv preprint arXiv:2010.08304 (2020)
7. Izhikevich, E.M., FitzHugh, R.: Fitzhugh-nagumo model. Scholarpedia **1**(9), 1349 (2006)
8. Jiang, X., Ghimire, S., Dhamala, J., Li, Z., Gyawali, P.K., Wang, L.: Learning geometry-dependent and physics-based inverse image reconstruction. in: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 487–496. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_47
9. Jiang, X., Missel, R., Toloubidokhti, M., Li, Z., Gharbia, O., Sapp, J.L., Wang, L.: Label-free physics-informed image sequence reconstruction with disentangled spatial-temporal modeling. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 361–371. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_35
10. Karl, M., Soelch, M., Bayer, J., Van der Smagt, P.: Deep variational bayes filters: unsupervised learning of state space models from raw data. arXiv preprint arXiv:1605.06432 (2016)
11. Kashtanova, V., Ayed, I., Cedilnik, N., Gallinari, P., Sermesant, M.: EP-Net 2.0: out-of-domain generalisation for deep learning models of cardiac electrophysiology. In: Ennis, D.B., Perotti, L.E., Wang, V.Y. (eds.) FIMH 2021. LNCS, vol. 12738, pp. 482–492. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78710-3_46
12. Krishnan, R.G., Shalit, U., Sontag, D.: Deep kalman filters (2015)
13. Li, L., Yan, J., Yang, X., Jin, Y.: Learning interpretable deep state space model for probabilistic time series forecasting (2021)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

15. Rangapuram, S.S., Seeger, M., Gasthaus, J., Stella, L., Wang, Y., Januschowski, T.: Deep state space models for time series forecasting. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 7796–7805 (2018)
16. Rubanova, Y., Chen, R.T.Q., Duvenaud, D.: Latent odes for irregularly-sampled time series (2019)
17. Smith, L.N.: Cyclical learning rates for training neural networks. In: 2017 IEEE Winter Conference on Applications Of Computer Vision (WACV), pp. 464–472. IEEE (2017)
18. Yildiz, C., Heinonen, M., Lahdesmaki, H.: Ode2vae: deep generative second order odes with bayesian neural networks. Adv. Neural Inform. Process. Syst. **32** (2019)
19. Yin, Y., et al.: Augmenting physical models with deep networks for complex dynamics forecasting. J. Statist. Mech. Theory Exper. **2021**(12), 124012 (2021)



Adaptive Unified Contrastive Learning for Imbalanced Classification

Cong Cong[1(✉)], Yixing Yang[1], Sidong Liu[2,3], Maurice Pagnucco[1], Antonio Di Ieva[3], Shlomo Berkovsky[2], and Yang Song[1]

[1] School of Computer Science and Engineering, University of New South Wales, Kensington, Australia
z3414050@ad.unsw.edu.au
[2] Centre for Health Informatics, Macquarie University, Sydney, Australia
[3] Computational NeuroSurgery Lab, Macquarie University, Sydney, Australia

Abstract. Medical image classifiers often suffer from the imbalanced class distribution of datasets. For example, among the 7 classes in the ISIC2018 skin lesion detection dataset, over 67% of the instances belong to melanocytic nevus while only 1% belong to dermatofibroma. Contrastive feature learning has been shown to achieve promising results in enhancing the performance for imbalanced classification tasks. However, the contrastive learning methods are either not end-to-end or require extra memory, which may lead to less compatible and sub-optimal features and classifiers. In this paper, we propose a novel unified feature and classifier learning framework for imbalanced medical image datasets. We equip our model with an adaptive unified contrastive (AduC) loss which progressively adapts model learning between feature learning and classifier learning. Furthermore, we explore the impact of different sampling methods on model training under data sparsity. The experimental results on two long-tailed medical datasets demonstrate that our methods can substantially improve the classification accuracy and F1-score over all classes without using extra memory storage. Our code is available at https://github.com/thomascong121/AdUni.

Keywords: Contrastive learning · Imbalanced classification

1 Introduction

Deep learning has shown great progress in medical image classification in recent years. However, many medical datasets are imbalanced, *i.e.*, majority classes have substantially more samples than other rare classes. As shown in Fig. 1, the majority classes (*e.g.*, "melanocytic nevus" in ISIC2018 [4] and "No Diabetic Retinopathy" in APTOS2019 [1]) dominate the other categories. Classification models that are trained under such imbalanced settings tend to be biased towards the majority classes and perform poorly on the minority classes.

To alleviate this issue in medical image datasets, data re-sampling techniques are widely used [2,12,25], which either up-sample the minority classes or down-sample the majority classes. Conditional generative networks [21] and the mix-up

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 348–357, 2022.
https://doi.org/10.1007/978-3-031-21014-3_36

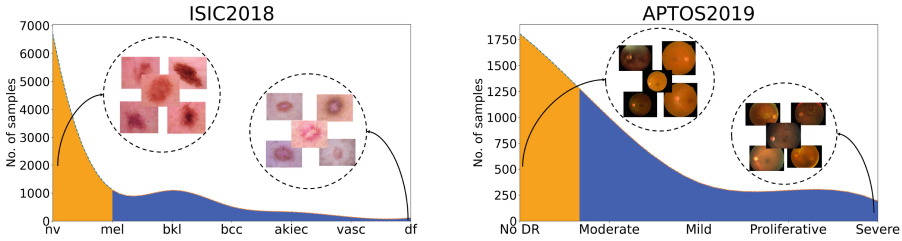


Fig. 1. Class distribution of the dataset used in this paper. Both datasets pose highly imbalanced distributions.

strategy [28] are also commonly used for generating synthetic samples [23,29]. Recently, Marrakchi *et al.* [17] adopted a supervised contrastive learning framework [11] and achieved substantially improved performance for lesion diagnosis and blindness detection under highly imbalanced settings. However, solutions for class imbalance in the medical domain are still under explored.

On the other hand, methods addressing class imbalanced learning for general classification tasks have been well studied. Re-sampling [19,24] and loss-sensitive learning [3,6,20] are the most widely used strategies. However, they normally gain improved performance on minority classes by sacrificing performance on majority classes and they ignore the different properties of feature learning and classifier learning [29]. Decoupled training [9,10,26,27,30] generally conducts feature learning in the first stage and performs classifier learning via cross-entropy (CE) in the second stage. Based on different feature learning methods, they can be divided into two categories: *classification pre-training* and *instance discrimination pre-training*. For classification pre-training [10,30], CE is typically used as the loss function which, however, has limited robustness against class imbalance (see Fig. 2 (left)). To alleviate this issue, these methods either use extra sampling methods (*e.g.*, square-root sampling [16]) or data mixup [28]. Contrastive learning (CL) is commonly used for instance discrimination pre-training, as it helps forming a more balanced feature space. For instance, Yang *et al.* [27] pre-trained the model using self-supervised contrastive learning; Kang *et al.* [9] proposed κ-positive contrastive learning for enhanced balancedness; and, Wang *et al.* [26] forced samples to be closer to their class prototype.

Although the decoupled training methods have demonstrated improved performance, these models are not trained end-to-end and are learned separately with two different targets for the feature learning and classifier learning stages, which makes the features learned from the first stage less compatible with the classifier in the second stage. To learn features and classifier compatibly, Parametric Contrastive Learning (PaCo) [5] provided a unified framework to perform both feature learning and classifier learning concurrently via a modified supervised contrastive loss and showed promising results. However, PaCo uses a memory queue [8] to store negative samples which impairs the class imbalance learning. To further illustrate this, we record the average number of instances of the majority class and minority class in a batch on APTOS2019 for PaCo and

our framework. On average, for each batch in our method, the number of majority class samples (70) is approximately 5 times greater than that of the minority class (14). On the other hand, since PaCo uses a memory queue of length 1024 (default value in [5]), the number of majority class samples (795) is over 10 times as many as the minority class samples (68). This indicates that using a memory queue exacerbates the biased learning towards the majority classes. Moreover, PaCo focuses more on classifier learning throughout the training. We argue such a biased learning is sub-optimal for medical datasets in which samples between classes share high semantic similarity (see Fig. 1).

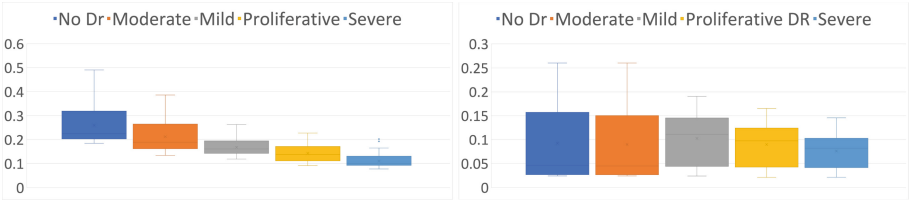


Fig. 2. We compare the loss magnitudes (y-axis) of 5 classes in the APTOS2019 dataset based on the model trained using cross-entropy loss (left) and AdUni (right). Classes are sorted by their cardinality. Clearly, AdUni shows a more balanced distribution of each class.

In this paper, we explore the effectiveness of feature learning in class imbalanced medical image classification. Since the memory queue mechanism in previous feature learning literatures may exacerbate the biased learning for imbalanced classification, we conduct both feature and classifier learning using batch contents only. Furthermore, inspired by the unified feature and classifier learning framework (PaCo), we propose an Adaptive Unified Contrastive (AduC) loss which well balances feature learning and classifier learning throughout training to obtain enhanced performance on datasets with high inter-class similarity. We also design an end-to-end unified learning framework (AdUni) which jointly performs feature learning and classifier learning using batch contents only with AduC. We conduct extensive experiments on the public ISIC2018 and APTOS2019 datasets, demonstrating that our method provides significant performance improvement over existing approaches.

2 Methodology

We design an Adaptive Unified learning framework (AdUni) (see Fig. 3) for imbalanced classification of medical images. AdUni, which is equipped with the proposed Adaptive Unified Contrastive (AduC) loss, provides a compatible learning process for feature and classifier without using any memory queue. As can be seen in Fig. 2, AdUni shows a relatively uniform loss distribution across classes which indicates that AdUni keeps a more balanced learning of each class. Both

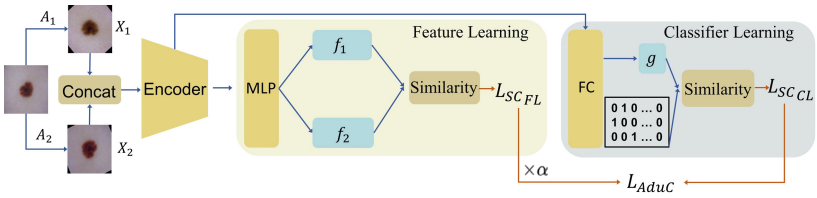


Fig. 3. The proposed AdUni framework which performs both feature learning and classifier learning via contrastive learning with Batch samples only. The model progressively transitions its learning focus from feature learning to classifier learning.

feature learning and classifier learning are conducted via supervised contrastive loss based on batch contents only and are jointly learned via our proposed adaptive unified mechanism to boost performance.

2.1 Feature Learning and Classifier Learning

Feature Learning via Contrastive Learning. We conduct feature learning using contrastive loss. Specifically, given a batch of N input images, for each image $x^i (1 \leq i \leq N)$, we obtain two augmentations x_1^i and x_2^i, respectively, which forms a total of $2N$ augmented training samples. Then, we obtain normalised feature vectors f_1^i and f_2^i using a ResNet-based encoder [7] followed by a two-layer multi-layer perceptron (MLP) (see Fig. 3). Finally, the classifier is trained with the features that are optimised by a supervised contrastive loss $\mathcal{L}_{SC_{FL}}$:

$$\mathcal{L}_{SC_{FL}} = -\sum_{p \in P(i)} \log \frac{\exp(f_i \cdot f_p)/\tau}{\sum_{j \in B(i)} \exp(f_i \cdot f_j)/\tau} \tag{1}$$

For each anchor sample with index i, $P(i)$ collects indices of positive samples (i.e., samples with the same label of sample i), $B(i)$ represents the other $2N-1$ samples and τ is the temperature parameter. Here, $\mathcal{L}_{SC_{FL}}$, which explicitly favours higher similarity between samples with the same class labels, pulls positive pairs closer to each other and forms denser clusters.

Classifier Learning via Contrastive Learning. Previous methods [10,17] use cross-entropy loss (CE) to perform classifier learning separately from feature learning. However, CE is not robust enough under class imbalance. Thus, we replace cross-entropy loss with contrastive loss which is shown to be more robust to class imbalance [9,13]. To achieve this, we use the form of supervised contrastive loss for classifier learning in [5] where we maximise similarities between the class logits $g \in \mathcal{R}^{2N \times C}$ produced by the encoder with the ground-truth label in one-hot format $l \in \mathcal{R}^{2N \times C}$, where C indicates the number of classes. This can be expressed as:

$$\mathcal{L}_{SC_{CL}} = -\sum_{p \in P(i)} \log \frac{\exp(g_i \cdot l_p)/\tau}{\sum_{j \in C} \exp(g_i \cdot l_j)/\tau} \tag{2}$$

where i indicates the index of anchor sample. $\mathcal{L}_{SC_{CL}}$ explicitly encourages the model to maximise the similarities between the logit g_i and the ground-truth label l_p over other classes.

2.2 Adaptive Unified Contrastive (AduC) Learning

To achieve joint feature and classifier learning, we first define a unified loss function L_{AduC}, which merges $\mathcal{L}_{SC_{FL}}$ and $\mathcal{L}_{SC_{CL}}$ by modifying the denominators:

$$\mathcal{L}_{AduC} = -\alpha \sum_{p\in\{P(i)\}} \log \frac{\exp(f_i \cdot f_p)/\tau}{\sum\limits_{j\in C} \exp(g_i \cdot l_j)/\tau + \sum\limits_{k\in B(i)} \exp(f_i \cdot f_k)/\tau} - \sum_{p\in\{P(i)\}} \log \frac{\exp(g_i \cdot l_p)/\tau}{\sum\limits_{j\in C} \exp(g_i \cdot l_j)/\tau + \sum\limits_{k\in B(i)} \exp(f_i \cdot f_k)/\tau}$$

The parameter α in Eq. 3 balances the loss function between feature learning and classifier learning. Generally, a smaller α weakens the influence of feature learning and strengthens the influence of classifier learning; and a larger α would make the model focus more on feature learning.

It has been observed that excessively focusing on classifier learning would lead to overfitting, whereas over-emphasising feature learning prevents the model from achieving a better accuracy. To capture the advantages of both feature learning and classifier learning and keep a balance between them, we propose to adaptively unify both training by gradually reducing α to smoothly transition model training from feature learning in the early stage of training to classifier learning in the later stage:

$$\alpha = \omega(t)\alpha_{max} + (1 - \omega(t))\alpha_{min} \tag{3}$$

Here, $\alpha_{min}/\alpha_{max}$ are the minimum/maximum values of α, t is the current epoch number and ω is a function which moderates the decrease of α.

3 Experiments and Results

3.1 Datasets and Implementation

We evaluated our methods on the ISIC2018 skin lesion detection dataset [4] and APTOS2019 blindness detection dataset [1]. ISIC2018 contains 10,015 images of size 450×600 pixels which are categorised into 7 disease states, including 6,705 melanocytic nevus (nv, 67%), 1,113 melanoma (mel, 11%), 1,099 benign keratosis (bkl, 11%), 514 basal cell carcinoma (bcc, 5%), 327 actinic keratosis (akiec, 3%), 142 vascular lesion (vasc, 1%) and 115 dermatofibroma (df, 1%). APTOS2019 contains 3,662 images with varying image sizes, which are categorised into 5 classes based on severity of diabetic retinopathy (DR): 1,805 (49%) no DR, 999 (27%) moderate, 370 (10%) mild, 295 (8%) proliferative, and 193 (5%) severe DR. Both the datasets pose data imbalance issues.

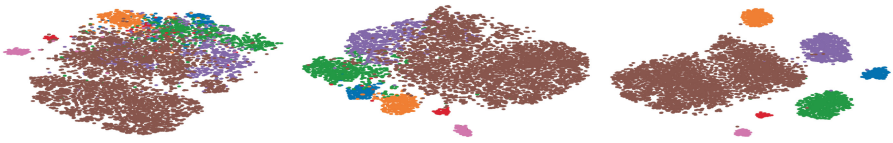
For both datasets, we conduct 5-fold cross validation and apply resize, random affine transformation, horizontal and vertical flip and color jittering for data augmentation. Furthermore, we resize images to 384×384 for training a ResNet18 backbone and evaluate its performance by measuring classification accuracy, averaged macro F1-score over all classes. For both datasets, we train the model for 400 epochs with batch size of 128 on 4 NVIDIA RTX3090 GPUs. We develop the framework using PyTorch [22] and set the initial learning rate to 0.1, which is decayed to 0.001 using the cosine schedule [14]. Training takes 5.5 h and 6.5 h on APTOS2019 and ISIC2018, respectively. We set $\tau = 0.1$, $\alpha_{min} = 0.01$, $\alpha_{max} = 1.0$, $\omega(t) = 1 + cos(t/t_{decay}\pi)$ and $t_{decay} = 1000$. For fair comparison, we re-implement all the baseline models with the same setting.

3.2 Results

Table 1. Classification accuracy and macro F1-score on the test set on both datasets using different methods.

Methods	ISIC2018		APTOS2019	
	Acc	Macro F1	Acc	Macro F1
CE [18]	0.841 ± 5.37e–03	0.701±1.21e–03	0.813 ± 6.09e–03	0.603 ± 7.81e–03
BALMS [24]	0.862 ± 2.28e–03	0.756 ± 1.94e–03	0.826 ± 3.46e–03	0.649 ± 3.19e–03
LDAM [3]	0.851 ± 2.28e–03	0.728 ± 1.94e–03	0.815 ± 2.05e–03	0.636 ± 3.46e–03
Decouple [10]	0.861 ± 2.32e–04	0.738 ± 4.01e–03	0.822 ± 2.83e–03	0.661 ± 3.31e–03
CICL [17]	0.866 ± 7.39e–03	0.760 ± 9.71e–03	0.828 ± 2.05e–03	0.676 ± 1.24e–03
PaCo [5]	0.864 ± 2.33e–03	0.768 ± 4.49e–03	0.825 ± 2.00e–03	0.680 ± 3.31e–03
PaCo+AduC	0.869 ± 1.79e–03	0.773 ± 1.25e–03	0.828 ± 2.05e–03	0.687 ± 1.73e–03
AdUni w/o AduC	0.876 ± 2.65e–03	0.778 ± 1.94e–03	0.826 ± 2.44e–03	0.682 ± 1.63e–03
AdUni	**0.878** ± 4.36e–03	**0.805** ± 1.18e–03	**0.839** ± 4.92e–03	**0.695** ± 5.51e–03

Comparison to the State-of-the-Art Methods. In Table 1, we compare the AdUni-based framework to state-of-the-art imbalanced classification studies. The compared methods [3,10,17,24] cover the three categories mentioned in Sect. 1, i.e., re-sampling, loss-sensitive learning and decoupled training, and a recent proposed unified method, i.e., PaCo [5]. Besides, we conduct PaCo+AduC in which we simply change the original fixed α with an adaptively changing α in AduC. It can be seen that our method consistently achieves the highest classification accuracy and macro F1-score.



(a) Decoupled [17] Features. (b) PaCo [5] Features. (c) AdUni Features.

Fig. 4. Feature visualisation of training images on ISIC2018 using t-SNE [15]

Among all the compared methods, the model trained with cross-entropy loss (CE) [18] has the lowest performance which indicates the limitation of CE for imbalanced classification. Decoupled training [10,17,24] produces comparable results; however, their learning schema are not end-to-end and the features learned from the first stage may become less compatible with the second stage learning. The unified training framework, PaCo [5], shows improved performance over the decoupled methods. PaCo+AduC can further enhance the performance of PaCo on both datasets which indicates the usefulness of AduC on datasets with similar semantic information across categories. Furthermore, we observe improved performance over PaCo even by removing the AduC loss from AdUni (AdUni w/o AduC) which illustrates the benefit of removing the memory queue. Moreover, as shown in Fig. 4, compared to the decoupled training method [17] which demonstrates superior performance on both datasets, the features learned by unified frameworks (PaCo [5] and AdUni are more visually separable. The proposed AdUni framework obtains a more distinguishable feature space than PaCo without using a memory queue. Furthermore, compared to the other methods, our proposed framework not only has the best overall classification performance but also achieves improved F1-score on every individual class for both the datasets, as shown in Fig. 5, demonstrating the consistent effectiveness of our method for both majority and minority classes.

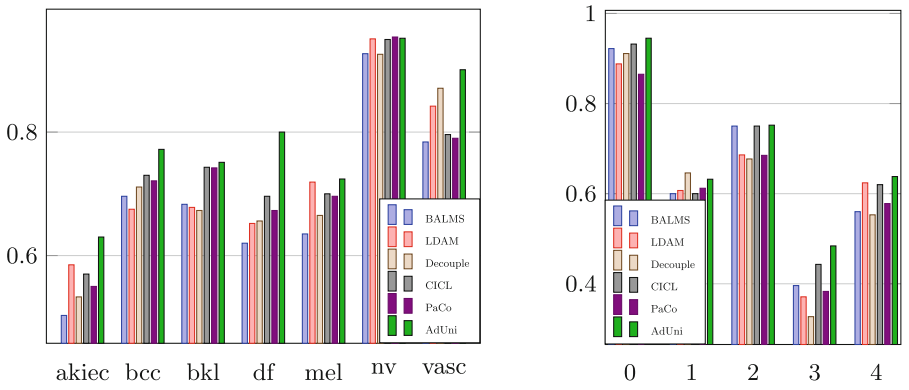


Fig. 5. F1-score for each class on the test set on ISIC2018 (left) and APTOS2019 (right) using different class-imbalanced studies.

Ablation Studies and Discussion. For ablation studies, we study the importance of different α values and show that our designed progressive adaptive α shows improved performance. Finally, we analyse the usefulness of data sampling for further performance improvement.

Table 2 presents the classification accuracy and macro F1-score using different α values and different sampling approaches. Unlike natural images, medical images typically have similar semantic information across categories; thus, we

Table 2. Classification accuracy and macro F1-score on the test set on both datasets with different sampling approaches.

		ISIC2018		APSTOS2019	
		Acc	Macro F1	Acc	Macro F1
$\alpha_{0.01}$	w/o sampling	0.867 ± 4.49e–03	0.760 ± 2.16e–03	0.822 ± 2.44e–03	0.675 ± 1.63e–03
	up-sampling	**0.876** ± 3.55e–03	**0.782** ± 5.79e–03	**0.830** ± 1.34e–03	**0.684** ± 3.74e–03
	SqRt sampling [16]	0.869 ± 3.29e–03	0.769 ± 2.94e–03	0.789 ± 3.87e–03	0.643 ± 1.89e–03
	ProBal sampling [3]	0.873 ± 1.70e–03	0.767 ± 4.77e–03	0.787 ± 6.48e–03	0.641 ± 6.02e–03
$\alpha_{1\to0.01}$	w/o sampling	0.878 ± 4.36e–03	0.805 ± 1.18e–03	0.839 ± 4.92e–03	0.695 ± 5.51e–03
	up-sampling	**0.886** ± 2.87e–03	**0.808** ± 2.78e–03	**0.841** ± 2.78e–03	**0.711** ± 3.54e–03
	SqRt sampling [16]	0.877 ± 3.19e–03	0.776 ± 1.72e–03	0.823 ± 2.49e–03	0.676 ± 4.90e–03
	ProBal sampling [3]	0.880 ± 1.72e–03	0.793 ± 3.27e–03	0.810 ± 3.99e–03	0.682 ± 3.26e–03

believe that a more separable feature space can make classifier learning easier. One way to focus the model on learning separable features is to use a large α throughout the training. However, we experimented with $\alpha = \{1, 100\}$ and found that setting large α normally leads to slow convergence. Thus, it is important to balance both feature learning and classifier learning. Here, we choose to adaptively change α using a cosine updating function $\omega(t) = 1 + cos(t/t_{decay}\pi)$ based on epochs t. By choosing suitable t_{decay}, we can keep α large in early training, then gradually drop it to smaller values in the later stage. We observe that using $\alpha_{max} = 1.0$ and progressively reducing it to 0.01 offers a reasonable choice. Table 2 shows that such mechanism improves the accuracy and F1-score by 2∼3%.

Moreover, we compare the results with different sampling approaches where up-sampling means we re-sample the minority classes to ensure every class has the same number of instances in each batch. Square-root (SqRt) sampling [16] calculates the sampling probability of each class using the square root of its sample numbers and progressive-balanced (ProBal) sampling [3] progressively increases the probability of minority samples until each class has equal chances to be selected. The results in Table 2 also show an interesting finding that the simple up-sampling approach works the best with AdUni and provides further improvement. We posit this may be due to the fact that contrastive learning-based approaches are more data-demanding. Although advanced sampling techniques like SqRt and ProBal increase the chance of minority classes to be selected, they cannot provide additional training data.

4 Conclusions

In this paper, we attempt to address the imbalanced data distribution issue for medical image classification. A novel end-to-end training framework is proposed, which jointly performs feature learning and classifier learning based on supervised contrastive loss and an adaptively unified mechanism. Our framework, while not requiring extra memory queue, improves classification performance for all classes in two public datasets ISIC2018 and APTOS2019. Our findings

highlight the benefits of enhanced feature learning in improving classification performance for medical image datasets which have an imbalanced data distribution and high similarity between classes.

References

1. APTOS 2019 blindness detection (2019). www.kaggle.com/c/aptos2019-blindness-detection/data
2. Bokhorst, J.M., Pinckaers, H., van Zwam, P., Nagtegaal, I., van der Laak, J., Ciompi, F.: Learning from sparsely annotated data for semantic segmentation in histopathology images. In: MIDL (2018)
3. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Adv. Neural Inform. Process. Syst. **32** (2019)
4. Codella, N., et al.: Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2019)
5. Cui, J., Zhong, Z., Liu, S., Yu, B., Jia, J.: Parametric contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 715–724 (2021)
6. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR, pp. 9268–9277 (2019)
7. He, K., Zhang, X., et al.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR, pp. 9729–9738 (2020)
9. Kang, B., Li, Y., Xie, S., Yuan, Z., Feng, J.: Exploring balanced feature spaces for representation learning. In: International Conference on Learning Representations (2020)
10. Kang, B., et al.: Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217 (2019)
11. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Adv. Neural Inform. Process. Syst. **33**, 18661–18673 (2020)
12. Li, Y., Shen, L.: Skin lesion analysis towards melanoma detection using deep learning network. Sensors **18**(2), 556 (2018)
13. Liu, H., HaoChen, J.Z., Gaidon, A., Ma, T.: Self-supervised learning is more robust to dataset imbalance. arXiv preprint arXiv:2110.05025 (2021)
14. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
15. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
16. Mahajan, D., et al.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 181–196 (2018)
17. Marrakchi, Y., Makansi, O., Brox, T.: Fighting class imbalance with contrastive learning. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 466–476. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_44
18. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT press (2012)

19. Park, S., Hong, Y., Heo, B., Yun, S., Choi, J.Y.: The majority can help the minority: context-rich minority oversampling for long-tailed classification. arXiv preprint arXiv:2112.00412 (2021)
20. Park, S., Lim, J., Jeon, Y., Choi, J.Y.: Influence-balanced loss for imbalanced visual classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 735–744 (2021)
21. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2337–2346 (2019)
22. Paszke, A., Gross, S., Massa, F., et al.: Pytorch: an imperative style, high-performance deep learning library. In: NeurIPS, pp. 8024–8035. Curran Associates, Inc. (2019)
23. Qasim, A.B., et al.: Red-GAN: attacking class imbalance via conditioned generation. yet another medical imaging perspective. In: Medical Imaging with Deep Learning, pp. 655–668. PMLR (2020)
24. Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al.: Balanced meta-softmax for long-tailed visual recognition. Adv. Neural Inform. Process. Syst. **33**, 4175–4186 (2020)
25. Reza, M.S., Ma, J.: Imbalanced histopathological breast cancer image classification with convolutional neural network. In: ICSP, pp. 619–624 (2018)
26. Wang, P., Han, K., Wei, X.S., Zhang, L., Wang, L.: Contrastive learning based hybrid networks for long-tailed image classification. In: CVPR, pp. 943–952 (2021)
27. Yang, Y., Xu, Z.: Rethinking the value of labels for improving class-imbalanced learning. Adv. Neural Inform. Process. Syst. **33**, 19290–19301 (2020)
28. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: beyond Empirical Risk Minimization. In: International Conference on Learning Representations (2018)
29. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: a survey. arXiv preprint arXiv:2110.04596 (2021)
30. Zhong, Z., Cui, J., Liu, S., Jia, J.: Improving calibration for long-tailed recognition. In: CVPR, pp. 16489–16498 (2021)



Prediction of HPV-Associated Genetic Diversity for Squamous Cell Carcinoma of Head and Neck Cancer Based on ^{18}F-FDG PET/CT

Yuqi Fang[1], Jorge Daniel Oldan[2], Weili Lin[1], Travis Parke Schrank[3,4], Wendell Gray Yarbrough[3,4,5], Natalia Isaeva[3,4,5], and Mingxia Liu[1(✉)]

[1] Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
mxliu@med.unc.edu

[2] Department of Radiology, Division of Nuclear Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[3] Department of Otolaryngology-Head and Neck Surgery, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[4] Linberger Comprehensive Cancer Center, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[5] Department of Pathology and Lab Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Abstract. The relationship of human papillomavirus (HPV) to squamous cell carcinoma of the head and neck has been long known and explored. However, it has not been investigated whether cancer areas have the prognostic value of genetic diversity within HPV. Most existing studies in head and neck cancer analysis only involve a single imaging modality, *e.g.*, computed tomography (CT) or magnetic resonance imaging (MRI), which may not provide complementary and diverse information for prediction task. Recently, positron emission tomography with 2-deoxy-2-[fluorine-18]fluoro-D-glucose integrated with CT (^{18}F-FDG PET/CT) has become a powerful imaging tool. In this study, we integrate these two imaging modalities (*i.e.*, PET and CT) for HPV-associated genetic diversity prediction. Specifically, we design a deep 3D convolutional neural network (called PCNet) to learn PET and CT features in a data-driven manner, consisting of two branches (with each one corresponding to a specific data modality). The generated intermediate feature maps are further fed into a fully-connected layer for abstraction. Moreover, radiomic characteristics, which have been verified as a prognostic indicator in head and neck cancer, are concatenated with these data-driven deep learning features for final prediction. Experiments on 50 subjects demonstrate the effectiveness of our proposed PCNet. This is among the first attempts to explore the potential of PET/CT in dif-

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_37.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 358–366, 2022.
https://doi.org/10.1007/978-3-031-21014-3_37

ferentiating genetic diversity in patients with HPV-associated squamous cell carcinoma of the head and neck.

Keywords: Head and neck cancer · Far clade classification · ^{18}F-FDG PET/CT · Convolutional neural network

1 Introduction

Head and neck cancer includes a group of cancers of the mouth, nose, or throat, and the most common type of head and neck cancer is squamous cell carcinoma [1,2]. The relationship of human papillomaviru (HPV) to squamous cell carcinoma of the head and neck has been long known and explored, but whether the cancer areas have prognosis value of the genetic diversity within HPV has not been investigated yet. In our recent work [3], oncogenic HPV16 viruses can be divided into two clades based on the distance from the reference A1 sequence, with the more-mutated, or far clade, having relatively favorable recurrence-free survival. In this study, we aim to investigate whether there is some underlying association between phenotypic characteristic of cancer areas and HPV16-related genetic diversity (*i.e.*, far clade or near clade).

Most existing studies in head and neck cancer analysis only involve a single imaging modality, *e.g.*, computed tomography [4] or magnetic resonance imaging (MRI) [5,6]. For example, Hermans *et al.* [4] explore the value of CT-determined tumor perfusion as a predictive factor of local failure and cause-specific survival in head-and-neck cancer treated by radiotherapy. Razek *et al.* [5] assess arterial spin labelling and diffusion-weighted MRI in differentiating recurrent head and neck cancer from post-radiation changes. Cardenas *et al.* [6] find that T2-weighted MRI can be used to evaluate the accuracy of head and neck normal tissue auto-segmentation systems through comparisons to expert manual labeling segmentations. However, single-modality setting in these studies cannot exploit complementary and diverse information conveyed in multi-modality data, leading to suboptimal performance.

Recently, positron emission tomography with 2-deoxy-2-[fluorine-18]fluoro-D-glucose integrated with CT (^{18}F-FDG PET/CT) has emerged as a powerful imaging tool. But only a few studies [7] use PET/CT in HPV16-associated head and neck cancer research. In this work, we aim to investigate whether PET/CT that incorporates metabolic activity can be used to differentiate genetic diversity in patients with HPV-associated squamous cell carcinoma of the head and neck. To this end, we design a deep 3D convolutional neural network (called **PCNet**) with two branches, each with PET and CT data as input, and their intermediate feature maps are concatenated, followed by a fully-connected layer for prediction.

Moreover, radiomic characteristics (*e.g.*, first-order statistics, shape, textural features) have been verified as a prognostic indicator in head and neck cancer analysis [8–10], including prediction of HPV status, differentiation between squamous cell cancer and benign tonsillar tissue or other malignancies such as lymphoma. We hypothesize that the integration of deep PET/CT features and

radiomic characteristics can provide complementary information, thus improving prediction performance. Accordingly, we incorporate radiomic characteristics of the cancer regions into the PCNet for feature concatenation, abstraction, and prediction. We validate the effectiveness of the proposed PCNet on 50 subjects with HPV-associated squamous cell carcinoma of the head and neck, with a promising classification accuracy. This indicates that there exists association between cancer tissues and HPV16-related genetic diversity, and ^{18}F-FDG PET/CT can be used as an objective tool to investigate such relationship.

2 Materials and Methodology

2.1 Materials and Data Preprocessing

A total of 50 patients with HPV-associated squamous cell carcinoma of the head and neck are included in this study. The pretherapy PET and CT images are obtained using a Siemens Biograph mCT or Biograph TruePoint scanner. The PET images are obtained 75 min (for dedicated head and neck images used during this study) after injection of 8–15 mCi (using a weight-based protocol of .154 mCi/kg) of F-18 FDG using an OSEM reconstruction with 4 iterations and 21 subsets, a 336×336 matrix, and 5 min per bed position. Each CT scan is obtained using 120 kVp, a 512×512 matrix, a weight-modulated mAs, and a 3 mm slice thickness.

The cancer regions of all subjects are manually segmented by an expert radiologist based on a clinical visualization software MIM (MIM Software, Cleveland, OH). Among the 50 subjects, 25 of them belong to *far clade* group (label: 1), while the remaining 25 subjects belong to *near clade* group (label: 0). We observe that the size of cancer regions of different subjects varies a lot. For instance, the cancer size of CT images varies from (5, 6, 3) to (57, 61, 24), where three elements in (:, :, :) denote the height, width, depth of the cancer regions. In this work, we resize the cancer areas of PET and CT images to (8, 8, 12) and (22, 20, 14), respectively. Additionally, we perform intensity normalization for these lesion regions using z-score standardization.

2.2 Methodology

Data-Driven Deep Learning Characteristics. In recent years, deep learning based on automatically learned task-oriented features has attracted researchers' interest in the field of medical image analysis [11–13], which facilitates us to learn PET/CT features in a data-driven manner. As demonstrated in Fig. 1, the proposed PCNet consists of a 3D CNN with two independent branches, with each branch designed to learn PET/CT-related features. The inputs of the two branches are 3D cancer regions with a size of (8, 8, 12) for PET data and (22, 20, 14) for CT data. In our experiments, we find that only two convolutional layers (*i.e.*, $Conv1$ and $Conv2$) are sufficient to extract features, which may due to the small cancer size. The $Conv1$ layer is with the input channel of 1 and the

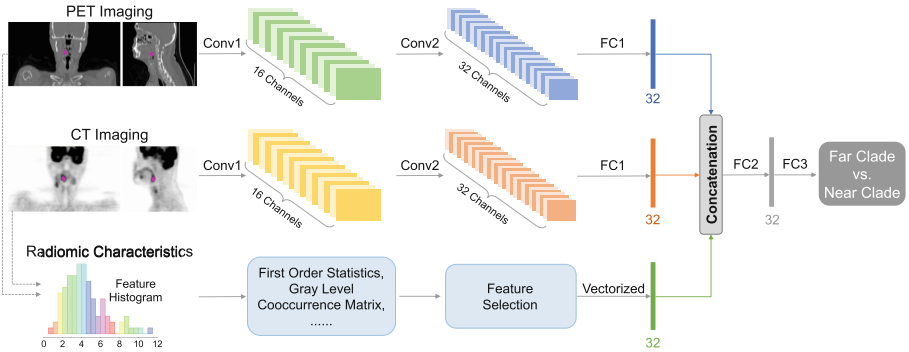


Fig. 1. Illustration of proposed PCNet in genetic diversity differentiation (far clade vs. near clade classification) in patients with HPV-associated squamous cell carcinoma of the head and neck based on PET and CT images.

output channel of 16, while the *Conv*2 layer has the input channel of 16 and the output channel of 32. Both *Conv*1 and *Conv*2 are with a kernel size of (3, 3, 3). Between two convolutional blocks, batch normalization (BN) [14] is used for feature standardization. The output of the fully-connected layer (*i.e.*, FC1) in each branch is a 32-channel feature vector representing input PET/CT data.

Radiomic Characteristics. Radiomic characteristics have demonstrated great clinical values in analysis of head and neck cancers, so we use such kind of features in this clade classification task. Specifically, 3D Slicer [15] is leveraged to derive 107-dimensional radiomic features of cancer areas for each subject from PET and CT images, including first-order statistics, shape, gray level co-occurrence matrix features, *etc.* All the 107 radiomic features are listed in Table SI of the *Supplementary Materials*. Since these radiomic features may be redundant for clade classification, we select the top 32 task-relevant features according to their Laplacian score in an unsupervised manner (without requiring any category label information) [16], which reflects the locality preserving power of each feature. The selected features are then combined with data-driven deep learning-based features for further classification, which will be detailed in the following.

Integration of Deep Learning and Radiomic Characteristics. With the hypothesis that the combination of handcraft radiomic features and deep learning based features may improve the clade classification performance, we integrate these features and feed them into two consecutive fully-connected layers in PCNet, with the input/output feature dimension of 96/32, and 32/1, respectively. The output is then passed through a *Sigmoid* activation function to perform the final classification. The whole network is trained through a binary cross-entropy loss L, which is mathematically formulated as:

$$L = -\sum_{i=1}^{N} y_i log(p_i) + (1 - y_i) log(1 - p_i). \tag{1}$$

By minimizing L, we aim to penalize the classification error by calculating the sum of the negative logarithm of predicted probabilities of each input, where $p_i \in (0, 1)$ denotes the predicted probability of the i-th subject categorized as the far clade group, $y_i \in \{1, 0\}$ means the ground-truth category label (*i.e.*, 1: far clade, 0: near clade), and N represents the number of training subjects.

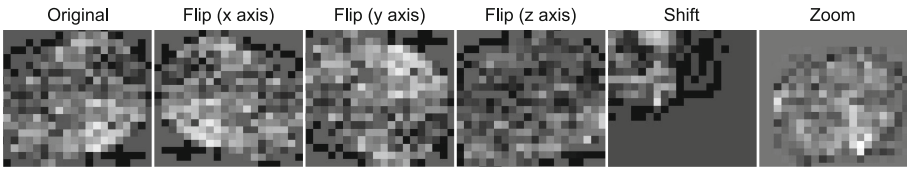


Fig. 2. Data augmentation methods performed on cancer areas of a CT image.

3 Experiment and Discussion

Competing Methods. To investigate the effectiveness of integration of PET and CT modality and the combination of radiomic and deep features, we compare our proposed method with the following seven methods. The **PET-R**, **CT-R**, and **PET&CT-R** methods only use radiomic features based on PET, CT, and PET&CT data, respectively, using an eXtreme Gradient Boosting (XGBoost) model [17] for classification. The **PET-D**, **CT-D**, and **PET&CT-D** methods only use deep learning based features to perform clade classification. The **PCNet-all** (a variant of our PCNet) method integrates both PET and CT modalities and abstracts their radiomic and deep learning based features to perform classification. The PCNet-all shares the same network architecture as PCNet. Note that PCNet uses only the selected 32-dimensional radiomic features based on their Laplacian scores, while PCNet-all leverages all radiomic characteristics for classification.

Experimental Setting. We divide 50 subjects into train and test set according to the split ratio of 7:3. Several 3D data augmentation methods are performed to increase sample size and improve model robustness, including flip (across x, y, and z axis), random shift, and random zoom (see Fig. 2). We ensure that the augmented images from the same subject are within either train or test set. The experiments are repeated for 100 times to reduce the bias caused by random data partition.

Six metrics are used for performance evaluation, including the area under the receiver operating characteristic curve (AUC), classification accuracy (ACC), F1 score (F1), sensitivity (SEN), specificity (SPE), and precision (PRE). We denote true positive (TP) as the number of subjects that are correctly classified as the positive group (*i.e.*, far clade); true negative (TN) as the number of subjects that are correctly classified as the negative group (*i.e.*, near clade); false positive

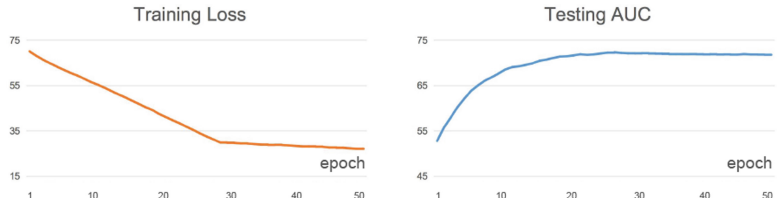


Fig. 3. The mean training loss and the testing AUC of the proposed PCNet.

Table 1. Results of the proposed PCNet in the task of far clade vs. near clade classification. Best results are shown in bold.

Method	AUC (%)	ACC (%)	F1 (%)	SEN (%)	SPE (%)	PRE (%)
PET-R	61.78 ± 13.11	57.44 ± 10.81	54.81 ± 13.84	54.37 ± 19.71	60.50 ± 20.13	59.26 ± 14.51
CT-R	62.28 ± 12.80	60.62 ± 11.47	58.93 ± 14.74	59.38 ± 19.56	61.88 ± 17.17	61.06 ± 13.09
PET&CT-R	67.33 ± 13.29	62.06 ± 11.67	60.59 ± 13.89	60.62 ± 18.74	63.50 ± 17.20	63.17 ± 13.89
PET-D	59.91 ± 12.69	57.19 ± 11.80	56.52 ± 13.18	57.38 ± 17.24	57.00 ± 18.48	58.01 ± 13.14
CT-D	67.47 ± 11.32	61.81 ± 10.60	61.69 ± 11.99	63.25 ± 16.94	60.38 ± 18.03	62.50 ± 12.04
PET&CT-D	67.45 ± 11.08	62.19 ± 10.09	61.05 ± 12.70	62.12 ± 18.99	62.25 ± 17.68	63.39 ± 11.95
PCNet-all	**73.41 ± 10.18**	**65.38 ± 9.62**	63.11 ± 11.80	61.00 ± 15.74	**69.75 ± 15.23**	**68.15 ± 13.07**
PCNet	71.81 ± 10.15	65.12 ± 10.50	**64.92 ± 11.79**	**66.38 ± 16.93**	63.88 ± 17.22	66.12 ± 12.75

(FP) as the number of subjects that are wrongly classified as the positive group; and false negative (FN) as the number of subjects that are wrongly classified as the negative group. Here, ACC $= \frac{TP+TN}{TP+FP+TN+FN}$, F1 $= \frac{2TP}{2TP+FP+FN}$, SEN $= \frac{TP}{TP+FN}$, SPE $= \frac{TN}{TN+FP}$, and PRE $= \frac{TP}{TP+FP}$.

Classification Results. The classification results of the competing methods are listed in Table 1. The training loss curve and testing AUC of our PCNet are shown in Fig. 3. From the table, we have the following observations. *First*, the integration of radiomic and data-driven deep learning features improves the classification performance by a large margin. Specifically, compared with PET&CT-R and PET&CT-D, our PCNet incorporating both characteristics achieves the improvement of 4.48% and 4.36% in terms of AUC values, respectively. This indicates that these features can provide complementary information for genetic diversity prediction. *Second*, methods trained with two modalities generally outperform their counterparts that are trained with a single modality. For instance, PET&CT-R surpasses PET-R and PET-D by an AUC of 5.55% and 5.05%, respectively, implying that diverse information conveyed in different modalities may help boost the classification results. *Moreover*, the methods using CT data show better classification results than those using only PET data. For example, CT-D outperforms PET-D by an improved AUC of 7.56%. The possible reason may be that CT imaging shows more accurate and detailed information for the organs and tissues compared with PET imaging, and CT images with higher resolution may help localize cancer areas more precisely.

Interpretation of Radiomic Characteristics. We use the Laplacian score algorithm to select the top 32 radiomic features from PET/CT with great locality preserving power. The *coarseness* feature ranks first among these features, which reflects the gray-level difference between the central voxel and its neighbors, indicating the spatial rate of change of image intensity. Many previous studies have also reported that coarseness serves as a significant predictor in cancer study [18–22]. For example, Yu *et al.* [18] extract textural features based on PET/CT and show that coarseness plays an important role in differentiating tumor from normal tissue in patients with head and neck cancer. Nakajo *et al.* [19] examine 40 pretreatment radiomic PET features in patients with endometrial cancer, and find that coarseness is an independent factor for both progression-free survival and overall survival. In our genetic diversity prediction task, the tumor coarseness reflects cancer aggressiveness that may be associated with clade-specific gene amplifications. Some other identified radiomics are also reported as important indicators in prognosis and disease grade/stage differentiation, including small dependence emphasis [23], small area low gray level emphasis [24], large area emphasis [25,26], and mean absolute deviation [27].

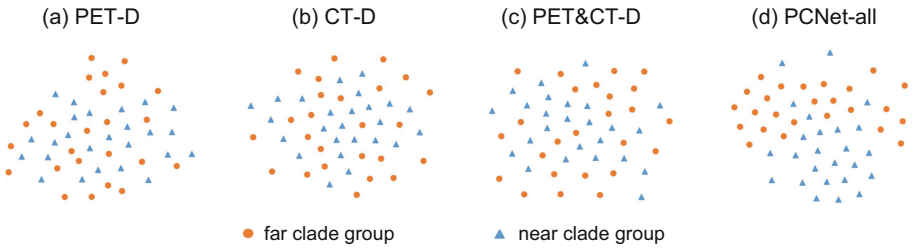


Fig. 4. T-SNE visualization of feature distributions of four deep learning methods, including (a) PET-D (b) CT-D (c) PET&CT-D, and (d) PCNet-all. Here, samples denoted by orange and blue are from far clade and near clade group, respectively. (Color figure online)

Visualization of Deep Learning Features. We further employ t-SNE [28] to visualize feature distributions of four deep learning methods (*i.e.*, PET-D, CT-D, PET&CT-D, and PCNet-all) in Fig. 4. For each method, we represent each subject using a 32-dimensional feature vector derived by the fully-connected layer FC1 (see Fig. 1). In Fig. 4, different colors represent different categories, *i.e.*, orange and blue denote the far clade and near clade groups, respectively. From Fig. 4, we can see that the inter-category data distribution gap based on features generated by our method is relatively larger than those by the other three methods, implying that our learned features are more discriminative for classification. This can partly explain why our method can produce better results than the competing methods in Table 1.

4 Conclusion

In this work, we design a deep 3D CNN incorporating radiomic and deep learning features from ^{18}F-FDG PET/CT data for HPV-associated genetic diversity prediction in patients with squamous cell carcinoma of the head and neck. Experiments on 50 subjects with ^{18}F-FDG PET and CT images demonstrate the effectiveness of our proposed method. Future direction will seek to validate the proposed method in large-scale studies, as well as to explore dynamic information on PET via list-mode acquisition and CT with multi-phase scanning.

References

1. Vokes, E.E., Weichselbaum, R.R., Lippman, S.M., Hong, W.K.: Head and neck cancer. N. Engl. J. Med. **328**(3), 184–194 (1993)
2. Syrjänen, S.: Human papillomavirus (HPV) in head and neck cancer. J. Clin. Virol. **32**, 59–66 (2005)
3. Schrank, T.P., et al.: Comprehensive viral genotyping reveals prognostic viral phylogenetic groups in HPV16-associated squamous cell carcinoma of the oropharynx. Molecular Cancer Research (2022)
4. Hermans, R., et al.: Tumor perfusion rate determined noninvasively by dynamic computed tomography predicts outcome in head-and-neck cancer after radiotherapy. Int. J. Radiat. Oncol. Biol. Phys. **57**(5), 1351–1356 (2003)
5. Razek, A.A.: Arterial spin labelling and diffusion-weighted magnetic resonance imaging in differentiation of recurrent head and neck cancer from post-radiation changes. J. Laryngol. Otol. **132**(10), 923–928 (2018)
6. Cardenas, C.E., et al.: Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations. Med. Phys. **47**(5), 2317–2322 (2020)
7. Aggarwal, C., et al.: Immunotherapy targeting HPV16/18 generates potent immune responses in HPV-associated head and neck cancer. Clin. Cancer Res. **25**(1), 110–124 (2019)
8. Wong, A.J., Kanwar, A., Mohamed, A.S., Fuller, C.D.: Radiomics in head and neck cancer: From exploration to application. Trans. Cancer Res. **5**(4), 371 (2016)
9. Jethanandani, A., et al.: Exploring applications of radiomics in magnetic resonance imaging of head and neck cancer: a systematic review. Front. Oncol. **8**, 131 (2018)
10. Fontaine, P., Acosta, O., Castelli, J., De Crevoisier, R., Müller, H., Depeursinge, A.: The importance of feature aggregation in radiomics: a head and neck cancer study. Sci. Rep. **10**(1), 1–11 (2020)
11. Fang, Y., Chen, C., Yuan, Y., Tong, K.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 302–310. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_34
12. Liu, M., Zhang, J., Lian, C., Shen, D.: Weakly supervised deep learning for brain disease prognosis using MRI and incomplete clinical scores. IEEE Trans. Cybern. **50**(7), 3381–3392 (2019)
13. Liu, M., Zhang, J., Adeli, E., Shen, D.: Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. IEEE Trans. Biomed. Eng. **66**(5), 1195–1206 (2018)

14. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. PMLR (2015)
15. Fedorov, A., et al.: 3D Slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging **30**(9), 1323–1341 (2012)
16. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. Advances in Neural Information Processing Systems 18 (2005)
17. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
18. Yu, H., Caldwell, C., Mah, K., Mozeg, D.: Co registered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning. IEEE Trans. Med. Imaging **28**(3), 374–383 (2008)
19. Nakajo, M., et al.: Application of a machine learning approach for the analysis of clinical and radiomic features of pretreatment [^{18}F]-FDG PET/CT to predict prognosis of patients with endometrial cancer. Mol. Imaging Bio. **23**(5), 756–765 (2021)
20. Lovinfosse, P.: FDG PET/CT radiomics for predicting the outcome of locally advanced rectal cancer. Eur. J. Nucl. Med. Mol.Imaging **45**(3), 365–375 (2018)
21. Zhong, J., et al.: Machine learning-based FDG PET-CT radiomics for outcome prediction in larynx and hypopharynx squamous cell carcinoma. Clin. Radiol. **76**(1), 78–e9 (2021)
22. Sellami, S., et al.: Predicting response to radiotherapy of head and neck squamous cell carcinoma using radiomics from cone-beam CT images. Acta Oncol. **61**(1), 73–80 (2022)
23. Liu, X., et al.: Computed tomography-based radiomics signature as a pretreatment predictor of progression-free survival in locally advanced hypopharyngeal carcinoma with a different response to induction chemotherapy. Eur. Arch. Otorhinolaryngology **279**(7), 3551–3562 (2022)
24. Yusufaly, T., et al.: Improved prognosis of treatment failure in cervical cancer with non-tumor positron emission tomography/computed tomography radiomics. J. Nucl. Med. **63**(7), 1087–1093 (2022)
25. Cook, G.J., Azad, G., Owczarczyk, K., Siddique, M., Goh, V.: Challenges and promises of PET radiomics. Int. J. Radiat. Oncol. Bio. Phys. **102**(4), 1083–1089 (2018)
26. Coroller, T.P., et al.: Radiographic prediction of meningioma grade by semantic and radiomic features. PloS One **12**(11), e0187908 (2017)
27. Kirienko, M., et al.: FDG PET/CT as theranostic imaging in diagnosis of non-small cell lung cancer. Frontiers Biosci. Landmark **22**(10), 1713–1723 (2017)
28. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11), pp. 2579-2605 (2008)



TransWS: Transformer-Based Weakly Supervised Histology Image Segmentation

Shaoteng Zhang[1,2], Jianpeng Zhang[2], and Yong Xia[1,2(✉)]

[1] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
[2] School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
yxia@nwpu.edu.cn

Abstract. Recently, weakly supervised histology image segmentation has received increasingly more attentions. Most solutions utilize a convolutional neural network (CNN) as a classifier and treat the generated class activation map (CAM) as a pseudo annotation, based on which a segmentation network is trained in a supervised manner. This pipeline suffers from two disadvantages. First, the CNN classifier may fail to generate the high-quality CAM that highlights the exact and integral target, resulting in incomplete activation and blurred boundaries. Second, it splits the original problem into two, leading to a sub-optimal solution and low efficiency. To address both issues, we propose a Transformer-based weakly supervised segmentation (TransWS) method for histology images. TransWS is composed of a classification branch and a segmentation branch. The former learns semantic information from image-level annotations and uses CAM to generate pseudo pixel-level annotations. The latter performs the class-agnostic segmentation (CAS), *i.e.*, binary segmentation, under the supervision of pseudo annotations. The semantic information and foreground region are combined to generate the final segmentation result. Comparing to CNN, Transformer is superior in modeling long-term dependencies and can generate more integral and accurate CAMs. More important, both branches in our TransWS can be jointly optimized in an end-to-end manner. We evaluated TransWS on the benchmark GlaS and Camelyon16-P512 datasets. Our results suggest that TransWS outperforms other weakly supervised segmentation competitors, setting a new state of the art.

Keywords: Weakly supervised segmentation · Histology images · Transformer · End-to-end optimization

1 Introduction

Histology image segmentation is crucial for computer-aided diagnosis. With the development of deep learning, convolutional neural networks (CNNs) have

S. Zhang and J. Zhang—Co-first authors.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 367–376, 2022.
https://doi.org/10.1007/978-3-031-21014-3_38

achieved remarkable success in image segmentation when sufficient annotations are available [3,12,22]. However, it is time-consuming and labor-intensive to annotate histology images on a pixel-by-pixel basis. Therefore, weakly supervised histology image segmentation has gained more and more attentions. Since [23], most weakly supervised segmentation methods have adopted a two-stage paradigm: (1) training a CNN as a classifier to extract the class activation map (CAM), which is further refined to form the pseudo pixel-level annotation, and (2) training a fully supervised segmentation model under the supervision of the pseudo annotation. However, these two-stage methods suffer from two disadvantages. First, since CNN usually has a limited ability to explore the global information, the CAM generated by a CNN can only highlight target regions roughly, resulting in incomplete activation and blurred boundaries. Second, these methods divide the weakly supervised segmentation task into two sequential sub-tasks, leading to sub-optimal solutions and less efficiency.

Recently, Belharbi *et al.* have argued that a CNN classifier may lack a mechanism to model the regions without a discriminatory power explicitly and to reduce false-positive rates [2]. Thus they proposed the regularization terms, which enable the model to seek both non-discriminative and discriminative regions, and achieved satisfactory results in the segmentation of colon cancer tissues and breast cancer tissues. This method, however, is still based on CNN and fails to utilize global information effectively. Qian *et al.* introduced Transformer to a multiple-instance learning (MIL) framework to capture global dependencies, and hence established the relationship among instances [13]. We argue that Transformer should be used to replace CNN in CAM based methods, aiming to improve the quality of CAMs and then generate reliable pseudo annotations.

In this paper, we propose a Transformer-based end-to-end weakly supervised segmentation (TransWS) method for histology images. TransWS is composed of a classification branch and a segmentation branch. Specifically, we use the image-level label to train the classification branch and treat the CAM generated by the classifier as pixel-level pseudo-labels. Then, the pseudo-labels are employed to train the segmentation branch. In addition, we also employ a class-agnostic segmentation (CAS) branch to further improve the segmentation performance. The final segmentation result is generated with the foreground region from the segmentation branch and semantic information from the classification branch. Comparative experiments demonstrate the superiority of our method, and our ablation studies suggest that each component we proposed improves the overall segmentation performance.

Our contributions are three-fold. First, we propose an end-to-end weakly supervised segmentation framework based on Transformer. Second, we improve the quality of CAMs by using Transformer. Third, our TransWS outperforms other competitors on two benchmark datasets, setting a new state of the art.

2 Method

Figure 1 shows the framework of TransWS. First, the input image is split into small patches and sent to the transformer encoder. Then, we convert the output

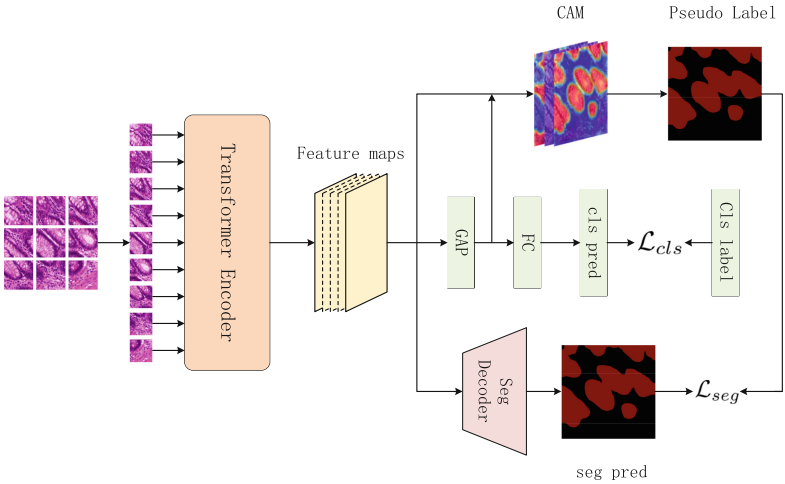


Fig. 1. Framework of the proposed TransWS.

of the transformer encoder into feature maps and input them to the classification and segmentation branches, respectively. We extract CAMs from the classification branch to generate pseudo-labels as the supervision of the segmentation branch. The classification loss is combined with our well-designed segmentation loss to optimize the entire network jointly. In addition, we use a class-agnostic segmentation(CAS) strategy to reduce the task difficulty of the segmentation branch. The final segmentation result is generated with the foreground region from the segmentation branch and semantic information from the classification branch.

2.1 Pseudo Label Generation

Given an image $I \in \mathbb{R}^{H\times W\times 3}$, we first split it into N patches, which are flattened and linearly projected to form N tokens $T_p \in \mathbb{R}^{N\times D_{in}}$, where D_{in} is the embedding dimension. The self-attention mechanism in Transformer is used to explore global feature dependencies. Specifically, patch tokens are linearly projected to form queries $Q \in \mathbb{R}^{N\times d_k}$, keys $K \in \mathbb{R}^{N\times d_k}$ and values $V \in \mathbb{R}^{N\times d_v}$, where d_k is the dimension of queries and keys, and d_v is the dimension of values. The Scaled Dot-Product Attention [20] is employed to compute the attention between queries and keys. The output token is formulated as:

$$X_i = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where X_i denotes the output of the i-th transformer encoder. After Transformer encoder, we get N patch tokens $X_{out} \in \mathbb{R}^{N\times D_{out}}$, where D_{out} denotes the embedding dimension of the last transformer encoder.

We generate CAMs in two steps [23]. First, We reshape the output token X_{out} into feature maps $F \in \mathbb{R}^{D_{out} \times \sqrt{N} \times \sqrt{N}}$. For a given class c, its activation map is formulated as

$$M^c = \sum_{i=1}^{D_{out}} W_i^c F_i, \tag{2}$$

where W^c denotes classification weights associated with the class c. Second, we employ the Min-Max normalization to limit the value of $M^c \in [0, 1]$, which is formulated as:

$$\hat{M^c} = \frac{M^c - min(M^c)}{max(M^c) - min(M^c)} \tag{3}$$

We denote all the foreground CAM as $M_{fg} = \text{Concat}(\hat{M^1}, \hat{M^2}, ..., \hat{M^C})$, where C is the number of foreground classes. Then we extract the background activation map from the following formula:

$$M_{bg}(i, j) = \{1 - \max_{c \in [0, C)} M_{fg}^c(i, j)\}^{\alpha}. \tag{4}$$

The segmentation pseudo label is defined as $M_s = \text{Concat}(M_{bg}, M_{fg}) \in [0, 1]$, which is upsampled to the original input resolution.

2.2 Class Agnostic Segmentation (CAS)

Since there is only one foreground class in each image for this study, we adopt the training strategy called Class Agnosic Segmentation (CAS) to differentiate the foreground from the background. The segmentation branch is trained under the supervision of M, which is formulated as follows:

$$M_h = argmax(M_s), M_h[M_h \neq 0] = 1, M = one_hot(M_h). \tag{5}$$

The final segmentation result is generated with the semantic information produced by the classification branch and the foreground region produced by the segmentation branch. This strategy significantly reduces the difficulty of the segmentation task and dramatically improves the performance.

2.3 Segmentation Loss Design

According to the pseudo segmentation label $M_s \in [0, 1]$, there are three possible cases for each label $M_c(i, j)$.
A. If $M_c(i, j)$ is close to 1, the pixel (i, j) is highly likely to belong to the class c.
B. If $M_c(i, j)$ is close to 0, the pixel (i, j) is not likely to belong to the class c.
C. Otherwise, it is hard to tell if the pixel (i, j) belongs to the class c.

Since the pseudo-label is produced by the classifier, the classifier is confident about the prediction at (i, j) in Case A and Case B, and is not confident enough in Case C. Therefore, we design the confidence weight, which takes a high value

when CAM is close to 0 or 1, and a low value when CAM is close to 0.5, shown as follows

$$W_{con}(i,j) = |\max_{c\in[0,C]} M_s^c(i,j) - 0.5| \times 2. \tag{6}$$

This design achieves our goal of focusing on optimizing the regions where the model is confident and suppressing the regions where the model is not confident.

In addition, we also observe that the distribution of foreground and background in the extracted pseudo-labels is unbalanced, so we also design another loss weight named balance weight, which is formulated as follows:

$$W_{ban}(i,j) = 1 - \frac{n}{N}, \tag{7}$$

where n represents the number of pixels in M_h with the same label to the pixel (i,j), and N refers to the total number of pixels in M_h. We denote the segmentation output as $S \in \mathbb{R}^{2\times H\times W}$. The segmentation loss is formulated as:

$$\mathcal{L}_{seg} = \sum_{i}^{H}\sum_{j}^{W}\left[\sum_{c=0}^{1} -M^c(i,j)\log(S^c(i,j)) \times W_{con}(i,j) \times W_{ban}(i,j)\right]. \tag{8}$$

2.4 Network Training

As shown in Fig. 1, our TransWS has two loss items, *i.e.*, the classification loss $\mathcal{L}_{cls}$ and segmentation loss $\mathcal{L}_{seg}$. Following the common practice, we feed the feature maps F to the classification layer and get the output P to compute the classification loss, shown as follows:

$$\mathcal{L}_{cls} = -Y \log P \tag{9}$$

where Y denotes the ground truth classification label, and P denotes the output probability of the classification branch. The overall loss is the sum of $\mathcal{L}_{cls}$ and $\mathcal{L}_{seg}$, which is formulated as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{seg} \tag{10}$$

3 Data

Two public datasets were used for this study. The GlaS dataset [17] contains 165 images of Hematoxylin and Eosin (H&E) stained slides for colon cancer diagnosis. For each image, both pixel-level and image-level annotations for cancer grading (i.e., benign or malign) are provided. The whole dataset is split into training (67 samples), validation (18 samples), and testing (80 samples) subsets [14].

The Camelyon16 dataset [1,14] has 399 whole slide images (both normal and metastatic) for metastases detection in H&E stained tissue sections of sentinel auxiliary lymph nodes (SNLs) in women with breast cancer. A protocol was developed to extract patches with image-level and pixel-level labels from this dataset [14]. For this study, we consider the benchmark containing patches of size 512×512 (referred to as Camelyon16-P512). This benchmark contains 48,870 samples, including 24,348 samples for training, 8,858 samples for validation, and 15,664 samples for testing [14]. All subsets have balanced global labels.

4 Experiments and Results

4.1 Evaluation Metrics

Following [2], we adopt the percentage of mis-classified samples (referred to as cls. error) to measure our model's classification performance, and the F_1 score over the foreground ($F_1^+ = \frac{2|\mathbb{G}^+ \cap \mathbb{S}^+|}{|\mathbb{G}^+|+|\mathbb{S}^+|}$) and the background ($F_1^- = \frac{2|\mathbb{G}^- \cap \mathbb{S}^-|}{|\mathbb{G}^-|+|\mathbb{S}^-|}$) to measure our model's segmentation performance, where $\mathbb{G}$ and $\mathbb{S}$ denote the pixel-level ground truth and segmentation result, respectively.

4.2 Implementation Details

We use the Mix Transformer (MiT) [18] as the backbone of our TransWS. Following [15], we set the stride to 1 in the last patch merging layer, aiming to keep the higher resolution for CAM. We utilize the multilayer perceptron (MLP) head [18] as the segmentation decoder. The Transformer backbone is initialized with the weights pre-trained on ImageNet [4].

We use the AdamW [9] algorithm to optimize our TransWS. The initial learning rate is set to 3×10^{-4} for GlaS and 1×10^{-4} for Camelyon16-P512, and the learning rate decays after each iteration with a polynomial scheduler. The maximum iterations are 500 and 20000 for GlaS and Camelyon16-P512, respectively. The weight decay factor is set as 0.01. To get better quality CAM, we warm up the classification branch for 200 and 2000 iterations for Glas and Camelyon16-P512, respectively, and save the model with the highest segmentation score ($F_1^+ + F_1^-$) on the validation set as our final model. The parameter α is set as 0.3 and 1.0 for GlaS and Camelyon16-P512, respectively. All experiments were conducted on a workstation with two NVIDIA Tesla P100 GPUs. The batch size is set to 6 per GPU.

4.3 Quantitative Results

Table 1 reports the classification and segmentation performance on the GlaS and Camelyon16-P512 test set. It can be seen that on the Glas dataset, our method surpasses all existing methods in both the classification and segmentation metrics. The F_1 score over the foreground and background is 2.12% higher than the second solution. On the Camelyon16-P512 dataset, although our F_1^- score is 2% lower than the first place, our method outperforms the existing methods in both classification performance and F_1^+ score. Shockingly, the F_1^+ score surpassed the second place by 15.34%. It is worth noting that the mean F_1 metric of our method far outperforms the other methods on both datasets.

To further demonstrate the effectiveness of our method, we implement experiments with end-to-end and two-stage methods based on CNN. The overall performance of the end-to-end method is better than that of the two-stage method, demonstrating the effectiveness of end-to-end optimization. Thanks to the Transformer backbone, the proposed TransWS outperforms the CNN counterparts by a large margin.

Table 1. Comparison of classification and segmentation performance on the GlaS and Camelyon16-P512 test set. mean F_1 denotes the average of F_1^+ and F_1^-. The best performance is shown in bold.

Dataset	GlaS				Camelyon16-P512			
Method	cls. error	F_1^+	F_1^-	mean F_1	cls. error	F_1^+	F_1^-	mean F_1
PN [8]	–	65.52	24.08	44.80	–	31.35	37.36	34.36
ERASE [21]	7.50	65.60	25.01	45.31	8.61	31.30	42.48	36.89
CAM-Max [10]	1.25	66.00	26.32	46.16	10.06	48.28	81.92	65.10
CAM-LSE [11,19]	1.25	66.05	27.93	46.99	1.51	64.31	63.78	64.05
Grad-CAM [16]	0.00	66.30	21.30	43.68	2.40	62.78	79.05	70.91
CAM-Avg [23]	0.00	66.90	17.88	42.39	2.40	62.75	79.05	70.90
Wildcat [5]	1.25	67.21	22.96	45.08	1.48	62.73	72.95	67.84
Deep MIL [7]	2.50	68.52	41.34	54.93	1.93	59.01	36.94	47.97
Max-Min Uncertainty(EEM) [2]	0.00	72.11	69.07	70.59	6.26	67.98	**88.80**	78.39
Max-Min Uncertainty(SEM) [2]	0.00	71.94	69.23	70.59	6.95	68.26	88.55	78.41
HistoSegResT [6]	0.00	74.04	71.16	72.60	–	–	–	–
CNN two-stage(stage 1)	**0.00**	69.69	45.54	57.61	**1.25**	75.49	78.76	77.13
CNN two-stage(stage 2)	–	50.53	69.52	60.03	–	34.49	71.30	52.89
CNN end-to-end	**0.00**	68.75	63.45	66.10	9.16	81.00	87.19	84.10
TransWS (ours)	**0.00**	**75.20**	**72.12**	**73.66**	1.28	**83.60**	86.80	**85.20**

4.4 Ablation Study

Table 2 shows the ablation experiment results on the GlaS test set. We observe that the proposed CAS, balance weight, and confidence weight are superior to boost the background and foreground segmentation quality, resulting in a great improvement for F_1^- and F_1^+. Compared with the baseline method in the first record, the proposed TransWS, integrating all of three components, obtains 20.94% and 38.42% performance gain of F_1^+ and F_1^-, respectively.

Table 2. Results of ablation experiments of our proposed method on the GlaS test set. ban_weight means balance weight and con_weight means confidence weight.

CAS	ban_weight	con_weight	F_1^+	F_1^-	mean F_1
×	×	×	54.26	33.70	43.98
✓	×	×	59.49	74.22	66.86
✓	✓	×	75.96	71.14	73.55
✓	×	✓	63.44	75.07	69.26
✓	✓	✓	75.20	72.12	73.66

4.5 Qualitative Results

Figure 2 visualizes the CAM and segmentation results of different methods obtained on the GlaS dataset. It reveals that (1) the CAMs and segmentation

results of CNN-based methods are not satisfactory; (2) the end-to-end method achieves clearer boundaries and more complete activation regions in CAMs and segmentation results than the two-stage method, and (3) the CAMs of our TransWS have more complete activation regions and more obvious boundaries, and our segmentation results are closer to the ground truth. It suggests that the end-to-end solution performs better than the two-stage one, and our TransWS is superior to CNN-based solutions with more accurate CAMs and segmentation predictions.

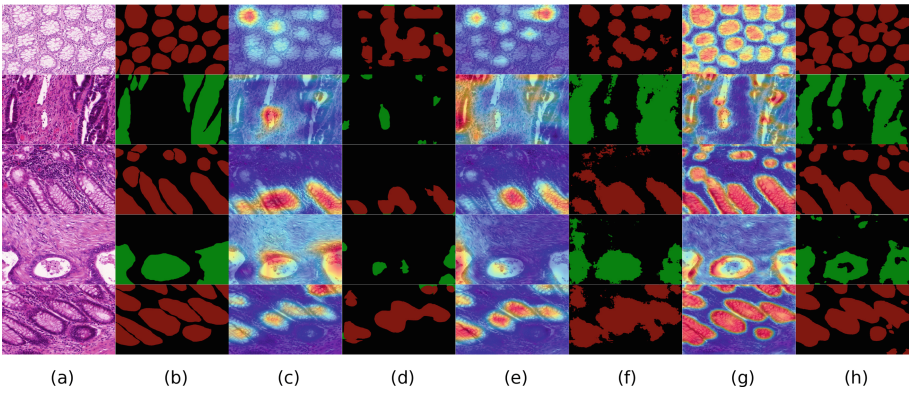


Fig. 2. Visualization of CAM and segmentation results on GlaS. (a) Image, (b) Ground truth, (c) CAM and (d) Segmentation result of two-stage CNN, (e) CAM and (f) Segmentation result of end-to-end CNN, (g) CAM and (h) Segmentation result of our TransWS, with red for benign and green for malignant.

5 Conclusion

In this paper, we propose a Transformer based end-to-end weakly supervised segmentation framework, called TransWS. Benefiting from the long-term modeling of Transformer, TransWS is able to generate accurate and complete pseudo-labels for the segmentation purpose. To achieve a joint optimal solution, both the classification and segmentation tasks are jointly optimized in an end-to-end manner. Extensive experiments demonstrate that TransWS outperforms all existing methods on the GlaS and Camelyon16-P512 benchmarks.

Acknowledgement. This work was supported in part by the Natural Science Foundation of Ningbo City, China, under Grant 2021J052, in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003, and in part by the Open Project of Ningbo Clinical Research Center for Medical Imaging under Grant 2022LYKFZD06.

References

1. Bejnordi, B.E., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph nodemetastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
2. Belharbi, S., Rony, J., Dolz, J., Ayed, I.B., McCaffrey, L., Granger, E.: Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty. IEEE Trans. Med. Imaging **41** (2021)
3. Chen, H., Qi, X., Yu, L., Heng, P.A.: Dcan: deep contour-aware networks for accurate gland segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2487–2496 (2016)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
5. Durand, T., Mordan, T., Thome, N., Cord, M.: Wildcat: weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 642–651 (2017)
6. Gu, W., Wang, S., Zhao, S., Wan, L., Zhu, Z.: Histosegrest: a weakly supervised learning method for histopathology image segmentation. In: 2022 the 5th International Conference on Image and Graphics Processing (ICIGP), pp. 189–195 (2022)
7. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
8. Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.B.: Constrained-CNN losses for weakly supervised segmentation. Med. Image Anal. **54**, 88–99 (2019)
9. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
10. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 685–694 (2015)
11. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1713–1721 (2015)
12. Qaiser, T., et al.: Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. Med. Image Anal. **55**, 1–14 (2019)
13. Qian, Z., et al.: Transformer based multiple instance learning for weakly supervised histopathology image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2022)
14. Rony, J., Belharbi, S., Dolz, J., Ayed, I.B., McCaffrey, L., Granger, E.: Deep weakly-supervised learning methods for classification and localization in histology images: a survey. arXiv preprint arXiv:1909.03354 (2019)
15. Ru, L., Zhan, Y., Yu, B., Du, B.: Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers. arXiv preprint arXiv:2203.02664 (2022)
16. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)

17. Sirinukunwattana, K., Snead, D.R., Rajpoot, N.M.: A stochastic polygons model for glandular structures in colon histology images. IEEE Trans. Med. Imaging **34**(11), 2366–2378 (2015)
18. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7262–7272 (2021)
19. Sun, C., Paluri, M., Collobert, R., Nevatia, R., Bourdev, L.: Pronet: learning to propose object-specific boxes for cascaded neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3485–3493 (2016)
20. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017)
21. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: a simple classification to semantic segmentation approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1568–1576 (2017)
22. Zhao, B., et al.: Triple u-net: hematoxylin-aware nuclei segmentation with progressive dense feature aggregation. Med. Image Anal. **65**, 101786 (2020)
23. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)



Contextual Attention Network: Transformer Meets U-Net

Reza Azad[1(✉)], Moein Heidari[2], Yuli Wu[1], and Dorit Merhof[1,3]

[1] Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany
{azad,yuli.wu,dorit.merhof}@lfb.rwth-aachen.de
[2] School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran
moein_heidari@elec.iust.ac.ir
[3] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany

Abstract. Convolutional neural networks (CNN) (e.g., UNet) have become the de facto standard and attained immense success in medical image segmentation. However, CNN based methods fail to build long-range dependencies and global context connections due to the limited receptive field of the convolution operation. Therefore, Transformer variants have been proposed for medical image segmentation tasks due to their innate capability of capturing long-range correlations through the attention mechanism. However, since Transformers are not designed to capture local information, object boundaries are not well preserved, especially in difficult segmentation scenarios with partly overlapping objects. To address this issue, we propose a contextual attention network that includes a boundary representation on top of the CNN and Transformer features. It utilizes an CNN encoder to capture local semantic information and includes a Transformer module to model the long-range contextual dependency. The object-level representation is included by extracting hierarchical features that are then fed to the contextual attention module to adaptively recalibrate the representation space using local information. In this way, informative regions are emphasized while taking into account the long-range contextual dependency derived by the Transformer module. The results show that our approach is amongst the top performing methods on the skin lesion segmentation benchmark, and specifically shows its strength on the SegPC challenge benchmark which also includes overlapping objects. Implementation code in github.

Keywords: Transformer · Semantic segmentation · Attention

1 Introduction

Convolutional neural networks (CNNs) have shown prominence in the majority of medical image segmentation applications. As a variant of these architectures, U-Net [23] has rendered notable performance and has been extensively utilized

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 377–386, 2022.
https://doi.org/10.1007/978-3-031-21014-3_39

across a wide range of medical domains [6,18,26]. In spite of their superb performance, CNN-based approaches suffer from a limitation in modeling the long-range semantic dependencies due to a confined receptive field size (even with dilated/atrous sampling [10]) and due to the nature of the convolution layer. Therefore, such a deficiency in capturing multi-scale information yields a performance degradation in the segmentation of complex structures, with variation in shapes and scales. Different methods have been proposed to solve the problem of the restricted receptive field of regular CNNs in recent years [10,20]. Wang et al. [27] extended the self-attention concept into the spatial domain to model non-local properties of images by devising a non-local module that can be easily integrated into existing network designs. As a result of the prominent performance of the attention mechanism, a line of researches has studied bridging the gap between the attention mechanism and CNNs in medical image segmentation [7,24]. To overcome the aforementioned limitation of CNNs, recently proposed Transformer-based architectures that leverage the self-attention mechanism to construct the contextual representations have been utilized. Transformers, unlike regular CNNs, are not only capable of modeling global contexts but also of leveraging to the versatile local information. Inspired by this, numerous studies have attempted to adapt Transformers for various image recognition tasks [8,13]. Moreover, Transformer-based models have recently gained growing attention ahead of their CNN counterpart in medical image segmentation [9,25]. As an example of an alternative perspective by treating semantic segmentation using Transformers, [29] proposed to model semantic segmentation as a sequence-to-sequence prediction task. However, the network is well designed to model the global contextual representation, it pays less attention to the local information and is consequently, less precise in the boundary area. Additionally, case studies have been established to investigate the amalgamation of Transformers and U-Net in medical image segmentation [9,15]. Valanarasu et al. [25] introduce a Local-Global training methodology for Transformers to learn both global and local features, respectively. However, this approach fails to model object-level interaction and renders a poor performance in the case of overlapping.

What all these methods have in common is their limitation in designing a specific mechanism to adaptively combine the global and local contextual representations. Particularly, a mechanism to jointly model the local semantic CNN representation along with the global contextual features derived from the Transformer module is critical for a task-specific purpose. To address this limitation, we propose the contextual attention network. Our design offers a two-stream pipeline, where in the first stream we utilize a CNN module to extract local semantic information and the object-level interaction map, while the second path incorporates the Transformer module to capture long-range contextual representations. Our Transformer module produces an image-level contextual representation (ICR) to construct the spatial dependency map in the image level and it produces regional importance coefficients (RIC) to model the importance of each region. In contrast to [25] which simply concatenates the local and global features, our method utilizes a contextual attention module to adaptively scale the feature maps and emphasizes on the important regions. Besides that we aim

to incorporate the object region information for precise boundary segmentation in case of overlapped objects. Through extensive experiments, our empirical findings validate that our method is able to pay more attention to the overlapped boundary area and while providing a strong semantic segmentation map. Our main contributions are summarized as 1: a contextual attention mechanism to adaptively aggregate pixel, object and image level features, 2: a coupling Transformer module with the CNN encoder to model object-level interaction.

2 Proposed Method

The proposed network architecture is depicted in Fig. 1. Our proposed architecture offers an end-to-end training strategy to adaptively incorporate the global contextual representation into local representative features derived from the CNN module. Our design proposes a contextual attention mechanism for feature recalibration and boundary-aware semantic segmentation. We will discuss each part in the following subsections.

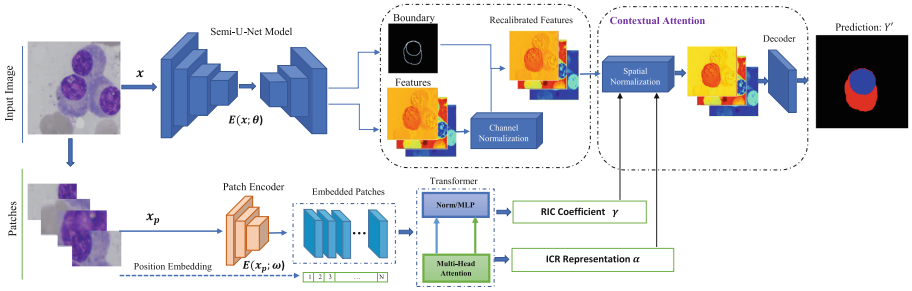


Fig. 1. Illustration of the proposed approach for medical image segmentation with the incorporation of the global contextual representation into local representative features.

2.1 CNN Representation

As illustrated in Fig. 1, our proposed method consists of two encoding streams, where in the first path we deploy the semi U-Net structure [23] to extract CNN representation. Given an image $\mathbf{x} \in \mathcal{R}^{H \times W \times C}$ with spatial dimension H and W, and C channels, our CNN encoder E_θ applies a series of convolutional blocks to model the pixel-level contextual representations. The locality nature of the convolutional operation usually limits the strength in modelling the object-level interaction. To include such representations, we model the object-level interaction by learning the boundary heatmap:

$$B = Conv_b(E(x;\theta)) \tag{1}$$

where the $Conv_b(.)$ shows a 1×1 kernel convolutional operation. This additional head enables the regional interaction and provides a surrogate signal for modelling the regional contextual dependency.

2.2 Long-range Contextual Representation

To learn the long-range contextual dependency we include the Transformer module in the bottom stream of our proposed pipeline. To prepare the input for the Transformer module, we divide the input image $\mathbf{x} \in \mathcal{R}^{H\times W\times C}$ into flattened uniform non-overlapping patches $\mathbf{x}_p \in \mathcal{R}^{N\times\left(p^2\cdot C\right)}$, where $p \times p$ denotes the dimension of each patch and $N = [\frac{HW}{p^2}]$ is the length of image sequence. Afterwards, using a patch encoder $E(x_p;\omega)$, we project the patches into a K dimensional embedding space. In order to maintain the spatial information of each patch, we learn a 1-D positional embedding $I_{\text{pos}} \in \mathcal{R}^{N\times K}$ which is subsequently added to the patch-embedding to preserve positional information $t_0 = \left[x_p^1 I; x_p^2 I; \cdots; x_p^N I\right] + I_{pos}$, where $I \in \mathcal{R}^{\left(p^2\cdot C\right)\times K}$ designates the projected patch embedding. We then exploit a stack of Transformer blocks encompassing a multi-head self-attention (MSA) and a multilayer perceptron (MLP) to learn the long-range contextual representation. An MSA layer is composed of M parallel self-attention heads to scale the embedded patches: $t_i' = \text{MSA}\left(\text{Norm}\left(\mathbf{t}_{i-1}\right)\right) + \mathbf{t}_{i-1}, \quad i = 1 \ldots L$. Next, the MLP modules learn the long-range contextual dependency by: $t_i = \text{MLP}\left(\text{Norm}\left(t_i'\right)\right) + t_i', \quad i = 1 \ldots L$, where Norm() denotes layer normalization and $t_i \in \mathcal{R}^{\frac{HW}{p^2}\times d}$ shows the encoded semantic representation in d dimensional space [4]. In addition to the encoded features, we model the image-level contextual representation *(ICR)* by reshaping *(Re)* the feature and applying a 1×1 convolutional operation ($Conv_I$):

$$ICR = Conv_I(Re(t_L)), ICR \in \mathcal{R}^{H\times W\times 1} \tag{2}$$

We use *ICR* to construct the spatial dependency map in the image level to later normalize the feature set generated by the CNN module. We further define the region importance coefficients *(RIC)* to model the distribution of foreground pixels in each region. The objective of the *RIC* coefficient is to provide a supervisory signal to guide the contextual attention module in determining the important regions (Eq. 3). $Conv_R$ shows a 1×1 convolutional operation.

$$RIC = Conv_R(t_L),\ RIC \in \mathcal{R}^{\frac{HW}{p^2}\times 1} \tag{3}$$

2.3 Contextual Attention Module

To adaptively aggregate the extracted features, we propose the contextual attention module. The importance of each feature set should be in line with the task at hand, thus, our proposed module utilizes the following two-level normalization steps: First, it recalibrates the CNN representation for a pixel-level object understanding. Then, it performs a spatial normalization to selectively emphasize the long-range contextual dependencies inside the feature set. Following SE module [17], we define the channel-wise normalization weights (w_{ch}) as:

$$w_{ch} = \sigma\left(\mathbf{W}_2\delta\left(\mathbf{W}_1 GAP(f)\right)\right) \tag{4}$$

where GAP shows the global average pooling operation applied to the CNN features (f), **W1** and **W2** are the learning parameters, and δ, σ are the ReLU and Sigmoid activation functions. We form the normalized features by: $f' = w_{ch} \cdot f$. To emphasize the boundary area, we add the boundary representation to each channel of the normalized feature, $\tilde{f} = f' + B$. The objective of the boundary feature is to emphasize the boundary regions and guide the model to precisely separate the overlapping objects (e.g. object-level interaction). Next, using the feature set derived from the Transformer module, we perform the spatial normalization. To this end, first we multiply the RIC coefficient with the corresponding regions (repeating the coefficient on the related region) in $\tilde{f}$ to scale the representation based on the regional importance, $f_{sn} = RIC \cdot \tilde{f}$. To further incorporate the long-range dependency, we concatenate the ICR representation with the f_{sn} and then apply the convolutional kernel followed by the batch normalization (BN) and activation to perform a non-linear aggregation:

$$\tilde{f}_{sn} = \delta(BN(Conv(ICR, f_{sn})) \tag{5}$$

The $Conv$ is the 1×1 convolutional operation. The resulting feature set contains both local semantic and global contextual representations which are selectively combined to perform the semantic segmentation task. Subsequently, we apply the decoder block to the extracted features to predict the segmentation mask, $Y' = D(\tilde{f}_{sn}; \gamma)$. The joint objective loss function that we optimize during the training is as follows:

$$\mathcal{L}_{\text{joint}} = \lambda_1 \mathcal{L}_{\text{segmentation}} + \lambda_2 \mathcal{L}_{\text{boundary}} + \lambda_3 \mathcal{L}_{\text{RIC}} \tag{6}$$

where $\mathcal{L}_{\text{segmentation}}$ calculates the Cross-entropy loss between the predicted mask and the ground truth, $\mathcal{L}_{\text{boundary}}$ shows the binary cross-entropy loss for the boundary prediction, and $\mathcal{L}_{\text{RIC}}$ calculates the MSE loss between the foreground pixel distribution in each image patch and the corresponding predicted one. We use coefficients $\lambda_i, i \in \{1, 2, 3\}$ to weight each loss.

3 Experiments

3.1 Dataset

Skin Lesion Segmentation: Three challenge benchmarks: ISIC 2017 [12], ISIC 2018 [11] and PH2 [21] are used in our evaluation. We followed [1,2] for data preparation and experimental setup.

Multiple Myeloma Segmentation: The proposed method is also evaluated on the multiple myeloma cell segmentation grand challenges, SegPC dataset [14]. Our Data preparation and evaluation setting follows [3].

3.2 Experimental Set-Up

Network Details and Training Process: As depicted in Fig. 1, our model uses both U-Net and Transformer modules to semantically label the input image.

For the U-Net model, we use the Resnet encoder [16] pre-trained on ImageNet and a four-blocks decoder module to generate the segmentation mask. Simultaneously, the Transformer structure follows the common implementation of the Vision Transformer with M (experimentally 4) heads. The implementation is performed in PyTorch and the results are carried out on a single GPU system with Nvidia RTX 3090. The model is trained end-to-end employing the Adam optimizer, batch size 4 and a learning rate 10^{-4} for 100 epochs.

Evaluation Protocol: Our evaluation takes into account the established evaluation metrics used in the respective challenges, which comprises sensitivity $= \frac{\text{TP}}{\text{TP+FN}}$, specificity $= \frac{\text{TN}}{\text{TN+FP}}$, accuracy $= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$, $mIOU = \frac{TP}{TP+FP+FN}$, and $DSC = \frac{2TP}{2TP+FP+FN}$ scores, where TP and TN indicates true and false samples which are correctly classified as true and false, respectively. FP and FN show the wrongly classified samples.

3.3 Results and Discussion

Quantitative Results: We provide qualitative comparisons on the benchmarks introduced in Sect. 3.1 and compare our method with the U-Net based literature work. Starting from the skin lesion segmentation scenario, Table 1 reports the comparison results on the three datasets, namely, ISIC 2017, ISIC 2018 and PH2. We exploited different evaluation metrics to accomplish a general and fair comparison. The baseline approach is simply the U-Net method without any of the proposed modules. Overall, our method attains a superior global performance across all datasets and evaluation metrics and most of the outperforming margins are statistically significant. Distinct improvements can be specifically observed opposed to pure CNN-based methods. We also observed that our method outperforms the Transformer [9,25,28] counterparts in almost all skin lesion segmentation benchmarks, which further proves the effectiveness of our design compared to other Transformer-based models. Specifically, compared to the FAT-Net, our method performs superior on both ISIC 2017 and ISIC 2018 datasets, which are large and challenging, but shows lower performance on the PH2 dataset (200 samples). The high number of parameters in our network might be a reason why our method did not perform well compared to the FAT-Net approach on the PH2 dataset. Additionally, Table 2 lists the quantitative results of different alternative methods [3,5] and the proposed network on the SegPC dataset. The results demonstrate that the proposed network attains better results than the other approaches by achieving significant performance gains over the baseline. It is worth mentioning that the SegPC dataset contains samples with high overlaps and the effectiveness of our method is statistically significant in comparison to the SOTA approaches. Our explicit modelling of the boundary information inside the network (opposite to the SOTA approaches) empowers the architecture to pay more attention to the overlapped objects. All in all, judging from the results in Table 1 and Table 2, one can see that our model attains excellent generalization on variable datasets, surpassing most of the CNN-based methods.

Qualitative Results: Visual segmentation results on both tasks are illustrated in Fig. 2 (also in the supplementary). Clearly, our proposed model

Table 1. Performance comparison of the proposed method vs. state-of-the-art methods on skin lesion segmentation benchmarks.

Articles	ISIC 2017				ISIC 2018				PH2			
	DSC	SE	SP	ACC	DSC	SE	SP	ACC	DSC	SE	SP	ACC
U-Net [23]	0.8159	0.8172	0.9680	0.9164	0.8545	0.8800	0.9697	0.9404	0.8936	0.9125	0.9588	0.9233
Att U-Net [22]	0.8082	0.7998	0.9776	0.9145	0.8566	0.8674	0.9863	0.9376	0.9003	0.9205	0.9640	0.9276
DAGAN [19]	0.8425	0.8363	0.9716	0.9304	0.8807	0.9072	0.9588	0.9324	0.9201	0.8320	0.9640	0.9425
TransUNet [9]	0.8123	0.8263	0.9577	0.9207	0.8499	0.8578	0.9653	0.9452	0.8840	0.9063	0.9427	0.9200
MCGU-Net [1]	0.8927	0.8502	**0.9855**	0.9570	0.8950	0.8480	**0.9860**	0.9550	0.9263	0.8322	0.9714	0.9537
MedT [25]	0.8037	0.8064	0.9546	0.9090	0.8389	0.8252	0.9637	0.9358	0.9122	0.8472	0.9657	0.9416
FAT-Net [28]	0.8500	0.8392	0.9725	0.9326	0.8903	**0.9100**	0.9699	0.9578	**0.9440**	**0.9441**	**0.9741**	**0.9703**
Proposed	**0.9158**	**0.9093**	0.9798	**0.9660**	**0.9059**	0.9038	0.9746	**0.9603**	0.9351	0.9382	0.9701	0.9605

produces smooth segmentation results for both tasks and performs well in the boundary area for separating the object of interest from the background. This fact reveals the importance of both Transformer modules for long-range contextual dependency (encouraging object learning) and the combination of boundary modules with the convolutional features maps in precise boundary recovery. Specifically, for the SegPC dataset, we observed that the proposed method segments the myeloma instances from a highly overlapped background with high precision. During our experimental visualization, we also observed that, compared to the U-Net model, the proposed structure produces robust segmentation results even with a noisy annotation, which is a common scenario in the medical domain. Comparable to recent work [9,25], our results also suggest that coupling the Transformer module with the CNN segmentation model can provide an additional input signal for a reliable and robust segmentation architecture. It should be noted that a drawback of our method compared to [3] is the high number of parameters (2 times higher) utilized in our network. Hence, depending on the application, one might consider the trade-off between the performance and computational complexity.

3.4 Ablation Study

The suggested network comprises the Transformer, the boundary and the attention add-on modules that are included for learning both local and global feature

Table 2. Performance evaluation on the SegPC challenge (best result is highlighted).

Methods	mIOU
FR U-Net [3]	0.9392
XLAB Insights [5]	0.9360
DSC-IITISM [5]	0.9356
MS Attention Deeplabv3+ [5]	0.9065
U-Net [23]	0.7665
Baseline	0.9172
Proposed	**0.9395**

Table 3. Effect of eliminating each module (**B**oundary, **T**ransformer or **C**ontextual **A**ttention) on the overall performance of the proposed method on ISIC 2018.

B	T	CA	DSC
	✓	✓	0.905
✓		✓	0.896
✓	✓		0.901
✓	✓	✓	**0.906**

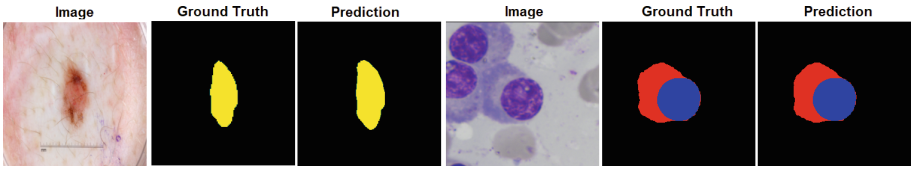


Fig. 2. Prediction results of the proposed method on both skin lesion segmentation and multiple Myeloma instance segmentation.

Table 4. Ablation study on the impact of the input size and model scaling. All experiments reported using the ISIC 2018 dataset.

Input-size influence			Model-scale influence		
Input-size resolution	DSC	ACC	Model structure	DSC	ACC
256 × 256	0.906	0.960	Base	0.906	0.960
512 × 512	0.918	0.975	Large	0.909	0.964

sets. To experimentally evaluate the effect and contribution of each module in the generalization performance, we selectively remove any of the modules, as shown in Table 3 (results on ISIC 2018). The qualitative finding suggests that removing any of the modules from the architecture results in a performance loss. More precisely, it can be observed that removing the Transformer module largely decreases the importance of the global attention maps and results in a clear performance drop. However, by including the Transformer module the network takes into account the strength of the global attention mechanism in combination with the local effectiveness of the convolution feature maps in learning generic and rich feature sets for precise localization ability. It is also worthwhile to mention that the combination of the proposed modules decreases the number of wrong predictions and the number of isolated FP (cf. Table 1) due to the strength of long-range contextual dependency maps. We further conducted ablation on the influence of input resolution and scaling the proposed method, both presented in Table 4 where we examined two different configurations, the "Base" and "Large" models. For model scaling we alter the Transformer number of layers, MLP size, and number of heads. Stating succinctly, we deduce that enlarging input resolution and model size leads to a performance boost, but the burden of computation cost, which guided us to adapt the "Base" model in our experiments.

4 Conclusion

In this work, we invoke the inherent frailty of regular convolutional neural networks in capturing long-range contextual dependencies. Specifically, the presented methodology is a novel contextual attention network which exploits the Transformer module along with the CNN encoder in order to concomitantly combine local and global representations for a further performance boost.

References

1. Asadi-Aghbolaghi, M., Azad, R., Fathy, M., Escalera, S.: Multi-level context gating of embedded collective knowledge for medical image segmentation. arXiv preprint arXiv:2003.05056 (2020)
2. Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S.: Bi-directional convlstm u-net with densely connected convolutions. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 406–415 (2019). https://doi.org/10.1109/ICCVW.2019.00052
3. Azad, R., Bozorgpour, A., Asadi-Aghbolaghi, M., Merhof, D., Escalera, S.: Deep frequency re-calibration u-net for medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3274–3283 (2021)
4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
5. Bozorgpour, A., Azad, R., Showkatian, E., Sulaiman, A.: Multi-scale regional attention deeplab3+: multiple myeloma plasma cells segmentation in microscopic images. arXiv preprint arXiv:2105.06238 (2021)
6. Cai, S., Tian, Y., Lui, H., Zeng, H., Wu, Y., Chen, G.: Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. Quant. Imaging Med. surg. **10**(6), 1275 (2020)
7. Cai, Y., Wang, Y.: Ma-unet: an improved version of unet based on multi-scale and attention mechanism for medical image segmentation. arXiv preprint arXiv:2012.10952 (2020)
8. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 357–366 (2021)
9. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
10. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
11. Codella, N., et al.: Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
12. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172. IEEE (2018)
13. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Gupta, A., Mallick, P., Sharma, O., Gupta, R., Duggal, R.: Pcseg: color model driven probabilistic multiphase level set based tool for plasma cell segmentation in multiple myeloma. PloS one **13**(12), e0207908 (2018)
15. Hatamizadeh, A., et al.: Unetr: transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
18. Huang, H., et al.: Unet 3+: a full-scale connected unet for medical image segmentation. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1055–1059. IEEE (2020)
19. Lei, B., et al.: Skin lesion segmentation via generative adversarial networks with dual discriminators. Med. Image Anal. **64**, 101716 (2020)
20. Li, M., Lian, F., Wang, C., Guo, S.: Accurate pancreas segmentation using multi-level pyramidal pooling residual u-net with adversarial mechanism. BMC Med. Imaging **21**(1), 1–8 (2021)
21. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J.: Ph 2-a dermoscopic image database for research and benchmarking. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5437–5440. IEEE (2013)
22. Oktay, O., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
24. Sinha, A., Dolz, J.: Multi-scale self-guided attention for medical image segmentation. IEEE J. Biomed. Health Inform. **25**(1), 121–130 (2020)
25. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: gated axial-attention for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 36–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_4
26. Valanarasu, J.M.J., Sindagi, V.A., Hacihaliloglu, I., Patel, V.M.: KiU-Net: towards accurate segmentation of biomedical images using over-complete representations. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 363–373. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_36
27. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
28. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: Fat-net: feature adaptive transformers for automated skin lesion segmentation. Med. Image Anal. **76**, 102327 (2022)
29. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)



Intelligent Masking: Deep Q-Learning for Context Encoding in Medical Image Analysis

Mojtaba Bahrami[1,2], Mahsa Ghorbani[1,2(✉)], Yousef Yeganeh[1], and Nassir Navab[1,3]

[1] Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany
[2] Sharif University of Technology, Tehran, Iran
mahsa.ghorbani@tum.de, mahsa.ghorbani@sharif.edu
[3] Whiting School of Engineering, Johns Hopkins University, Baltimore, USA

Abstract. The need for a large amount of labeled data in the supervised setting has led recent studies to utilize self-supervised learning to pre-train deep neural networks using unlabeled data. Many self-supervised training strategies have been investigated especially for medical datasets to leverage the information available in the much fewer unlabeled data. One of the fundamental strategies in image-based self-supervision is context prediction. In this approach, a model is trained to reconstruct the contents of an arbitrary missing region of an image based on its surroundings. However, the existing methods adopt a random and blind masking approach by focusing uniformly on all regions of the images. This approach results in a lot of unnecessary network updates that cause the model to forget the rich extracted features. In this work, we develop a novel self-supervised approach that occludes targeted regions to improve the pre-training procedure. To this end, we propose a reinforcement learning-based agent which learns to intelligently mask input images through deep Q-learning. We show that training the agent against the prediction model can significantly improve the semantic features extracted for downstream classification tasks. We perform our experiments on two public datasets for diagnosing breast cancer in the ultrasound images and detecting lower-grade glioma with MR images. In our experiments, we show that our novel masking strategy advances the learned features according to the performance on the classification task in terms of accuracy, macro F1, and AUROC.

Keywords: Self-supervised learning · Context encoding · Reinforcement learning · Medical image analysis

M. Bahrami and M. Ghorbani—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_40.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 387–396, 2022.
https://doi.org/10.1007/978-3-031-21014-3_40

1 Introduction

Deep supervised neural networks need a substantial amount of annotated data to demonstrate state-of-the-art performance, particularly in computer vision tasks. However, such an amount of labeled data is hard to acquire especially in medical imaging. Therefore, methods that do not require large data are becoming more interesting to train high-accuracy models with a fewer number of annotated samples. Self-supervision is a family of methods to pre-train deep neural networks using a large number of unlabeled data. The resulting pre-trained networks are then fine-tuned for downstream tasks with much fewer domain-specific labeled samples. Self-supervised strategies showed superior performance not only in learning natural images [7,8,11,15,16], but also for medical images [2,5,10,17,23,25] in various tasks including classification, segmentation, anomaly detection, and localization.

Recently, pre-training with pretext task has become more popular. Pretext task learning strategies extract auxiliary labels from unlabeled data and then solve a proxy task to learn semantic representations. Pathak et al. [16] proposed an in-painting strategy where part of the unlabeled image is masked and a reconstruction model is supposed to predict the masked area. Other methods were also proposed to use the relative positions of image patches [7], rotation [8], and puzzle-solving [15] as the self-supervision task.

Similar but more customized self-supervised strategies are also explored in the domain of medical image analysis. Jamaludin et al. [10] proposed the usage of follow-up MR image scans from the same patients in longitudinal studies as a free self-supervision task. They trained a Siamese CNN to recognize the scans that come from the same patients. In [5], the authors proposed a patch swapping strategy followed by a context restoration network to reconstruct the original image. Other methods include pre-training with 2D sliced image order prediction [23] for fine-grained body part recognition, recovering 3D medical images as a Rubik's cube [25], and predicting anatomical positions in MR images [2]. However, in the existing methods, the transformations applied in order to build a self-supervision dataset are generally blind with random masking, rotation, and swapping. We believe that choosing the appropriate transformations instead of random ones can improve both the efficiency and effectiveness of the training process. To this aim, we employ Deep Reinforcement Learning (DRL) to learn the optimal masking strategy. Recent works have demonstrated the great potential of DRL in medical applications [24] such as landmark detection [18,22], image registration [12], and image segmentation [13,21]. Yet none of these studies focuses on improving the pre-training task in self-supervised learning. In this work, we propose a novel context prediction framework that masks the most informative and effective parts of an image for our pretext task learning. We developed a Reinforcement Learning (RL) agent that learns the optimal masking strategy against a prediction model that tries to predict the masked region.

2 Methods

The main challenge behind finding an effective masking strategy is that choosing the mask area is a non-differentiable operation that cannot be easily optimized through an end-to-end gradient descent approach. The idea developed in this paper is to leverage the capability of reinforcement learning in optimizing the masking strategies to propose intelligent mask areas. Our method consists of two components, including a Q-learning-based masking agent and a prediction model that is pre-trained with our in-painting pretext task. As our masking model is based on deep Q-learning, we first give a brief description of the Q-learning algorithm and then present our main contributions.

2.1 Deep Q-Learning

Reinforcement learning (RL) is used to train optimal policies where there are no input-output pairs as in the supervised setting or when we need to optimize a policy over a discrete action space. Each environment is modeled by a set of states $s \in S$ that give the instant reward $R(s, a)$ by taking the action $a \in A$. A reinforcement setting is modeled by a finite Markov Decision Process (MDP) in which the transition to the next state s_{t+1} only depends on the current state s_t and the chosen action a_t. The dynamic of the environment is stated by the transition matrix $T(s, s', a)$ containing the probability that taking action a at state s will lead to state s' in the next step. The final goal is to find an optimal policy $\pi(s)$ for the agent that maximizes the total expected return $\mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$ where γ is the discount factor of the future rewards against the instant ones. However, in most real-world problems, the dynamics of the environment $T(s, s', a)$ are not available. Q-learning [20] is a model-free RL algorithm that does not need to explicitly know the dynamics of the environment. For this purpose a value function $Q(s, a) : S \times A \rightarrow \mathbb{R}$ is learned to approximate the expected rewards for all actions at each given state using the following iterative formula derived from the Bellman optimality equation:

$$Q_{new}(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \cdot \left[R(s_t, a_t) + \gamma \cdot \max_{a'} Q\left(s'_{t+1}, a'\right) - Q(s_t, a_t)\right] \quad (1)$$

where α is the learning rate of the agent. At each step, the agent corrects its approximation of the total expected future rewards. By updating the Q-function for a sufficient number of iterations the Q-function converges according to [19]. Consequently, the optimal policy for the agent can be inferred by taking the action that maximizes the Q-value function at each state. Furthermore, in deep Q-learning, the Q-value function is approximated by a deep neural network with the state s_t as the input and the Q-values as its output.

2.2 Self-supervised Intelligent Masking

Each self-supervision task is composed of two steps: 1) creating paired input-output images and 2) training a prediction model to learn the mapping. In the first step, for a given sample $\mathbf{x_i}$ from dataset $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$, we produce

an occluded variant $\tilde{\mathbf{x}}_\mathbf{i}$, and in the second step we train a prediction model to learn the inverse mapping of $\tilde{\mathbf{x}}_\mathbf{i} \rightarrow \mathbf{x}_\mathbf{i}$. As shown in Fig. 1, we introduce an intelligent masking mechanism to occlude a region based on the input image context to improve the semantic features learned by the prediction model and its performance on subsequent downstream tasks. In the following sections, we present the masking and prediction methods in detail.

Prediction Network: We include a prediction model that receives a transformed variant $\tilde{\mathbf{x}} = (1 - \hat{M}) \odot \mathbf{x}$ of an input image $\mathbf{x}$ where $\odot$ is an element-wise product operation and $\hat{M}$ is a binary mask occluding an arbitrary region of the input image. The prediction model is then pre-trained to reconstruct the masked region of the input image $\mathbf{x}$ using the context information. To this aim, we use a deep convolutional neural network with an Encoder-Decoder architecture $f(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathcal{D}_{\boldsymbol{\phi}}(\mathcal{E}_{\boldsymbol{\theta}}(\mathbf{x}))$ to capture the semantics of the input image. The encoder network $\mathcal{E}_{\boldsymbol{\theta}}$ consists of a series of stacked convolution blocks followed by pooling and batch normalization layers with parameters $\boldsymbol{\theta}$. The decoder network $\mathcal{D}_{\boldsymbol{\phi}}$ with parameters $\boldsymbol{\phi}$, includes a set of transposed convolution layers in order to increase the spatial dimension of the latent embedding to generate an output with the same size as the input. Inspired by [16], we also apply a channel-wise convolutional layer right after the encoder to spread the information through the spatial dimension from one corner to the other in the latent feature map. We use pixel-wise Mean Squared Error (MSE) as the loss between the output of the prediction network and the original missing region over the masked area. Therefore, the objective of the prediction model is defined as follows:

$$\min_{\boldsymbol{\theta}, \boldsymbol{\varphi}} \mathcal{L}_{pred}(\tilde{\mathbf{x}}, \mathbf{x}) = \|\hat{M} \odot [\mathbf{x} - \mathcal{D}_{\boldsymbol{\phi}}(\mathcal{E}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}))]\|_2 \tag{2}$$

where the MSE loss is only calculated over the masked region. As stated by [16], although the MSE loss may prefer a blurry reconstruction of the masked region; however, it is sufficient for the feature learning task.

Masking Network: We propose a novel strategy to improve the pretext self-supervision in-painting task by proposing targeted binary masks $\hat{M}$ instead of random occlusions. To this end, we design a mechanism to mask a region of the input image that the prediction network is weak in its reconstruction. However, selecting the mask area is a discrete and non-differentiable operation that cannot take advantage of gradient descent optimization. Therefore, we develop a masking framework based on reinforcement learning to learn the optimal masking policy. To develop such an agent, we design a 1-step episodic task in which the input image is the current state $s = \mathbf{x}$ that creates a continuous infinite state space. We also define the set of actions A to be the set of $k \times k$ overlapping image patches that can be selected to create the corresponding binary mask $\hat{M}$. This introduces a finite action space of size $|A| = k^2$ where k is a hyper-parameter. Finally, we set the instant reward for our agent to be the loss of the prediction model $R(s = \mathbf{x}, a) = \mathcal{L}_{pred}(\tilde{\mathbf{x}}, \mathbf{x})$. We employ Q-learning to train the agent by approximating a Q-value function with a convolutional neural network $Q(s, a; \psi)$ with parameters ψ. The network tries to predict the total

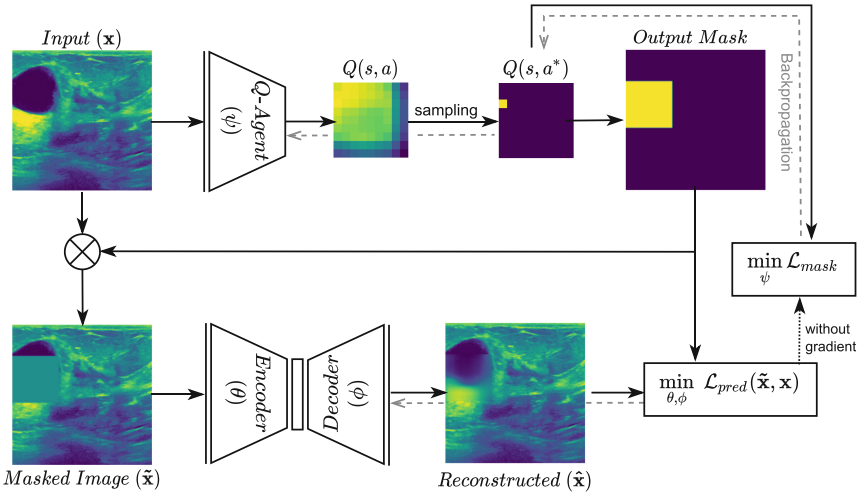


Fig. 1. Overview of the proposed model. In the masking network (upper part), the agent selects a patch of the image to be occluded based on its reward estimation of masking each patch. The original image is occluded with the mask created by the agent which is then predicted by the prediction network (lower part).

expected rewards for all possible masking actions given the input image. However, unlike the typical reinforcement settings, we only have single-step episodes. In each episode the agent receives the state information (input image $\mathbf{x}$), takes an action $a^* = \text{argmax}_a Q(s = \mathbf{x}, a)$ to mask an arbitrary region, and then the prediction model tries to reconstruct the masked region resulting in the loss value of $\mathcal{L}_{pred}$. After reconstruction, the agent receives the reward $R(s, a) = \mathcal{L}_{pred}$ and the episode terminates. For a single-step episode, the update rule of the Q-values becomes as follows:

$$Q_{new}(s, a) \leftarrow Q(s, a) + \alpha \cdot [R(s, a) - Q(s, a)] \tag{3}$$

where the $R(s, a)$ is set to $\mathcal{L}_{pred}$. In order to train our network $Q(s = \mathbf{x}, a; \boldsymbol{\psi})$ with the above update rule, we define the objective of the masking Deep Q-Network (DQN) according to [9] as:

$$\min_{\boldsymbol{\psi}} \mathcal{L}_{mask} = \left\| Q(s = \mathbf{x}, a^*, \boldsymbol{\psi}) - \mathcal{L}_{pred}(\tilde{\mathbf{x}}, \mathbf{x}) \right\|^2 \tag{4}$$

where a^* is the action selected to produce the masked variant $\tilde{\mathbf{x}}$. During our training, to have a balance between exploration and exploitation, we sample the actions based on the distribution of the action scores by applying a soft-max function over the Q-value action scores. Furthermore, to improve the stability of the training process, we adopt a two-phase training scheme where the masking agent operates completely random for the first few epochs.

3 Experiments and Results

In this section, we evaluate the transferability of our pre-trained network to the disease classification task, using the encoder $\mathcal{E}_{\theta}$ as a feature extractor. We experiment on two datasets: Low-Grade Gliomas (LGGs) [4,14] and Breast UltraSound Images (BUSI) dataset [1] in Sect. 3.1 and Sect. 3.2, respectively. In Sect. 3.3, we provide a qualitative example comparing other feature learning strategies. We compare our method with one unsupervised and two other self-supervised strategies. They include an unsupervised auto-encoder (*Reconstruction*) [3], context prediction (in-painting) with random masking (*Context Prediction*) [16], and self-supervision using disordering random patches and restoring the original image (*Context Restoration*) [5]. For both datasets, the whole training set is used for the self-supervision task. The performance of the trained classifier is evaluated in terms of accuracy, macro F1 score, and AUROC. We compare the performance of the pre-training phase in feature extraction especially when the training samples are limited.

Implementation Details: We start the training of both the self-supervised model and the classifier with the learning rate of 10^{-4}. A multi-step scheduler decays the learning rate by a factor of 0.3 in two equal milestones during both the self-supervised and classification training. We also include the same set of relatively strong augmentations for all methods to ensure that an exhaustive augmentation cannot achieve the effectiveness of a guided masking strategy. For this, we use random horizontal and vertical flip, rotation, and cropping for all methods. Besides, we utilize dropout with the rate of $p = 0.3$ just for fully-connected layers in the classifier and Adam optimizer [6] is selected for all trainings. More details of the implementation can be found in the supplementary materials.

3.1 Diagnosis of Lower-Grade Glioma

In our first experiment, we model the presence of the abnormality in the brain MR images as a binary classification with the purpose of Lower-Grade Glioma (LGG) detection. To this aim, we use Lower-Grade Glioma (LGG) dataset [4,14] obtained from The Cancer Imaging Archive (TCIA). The dataset contains 3929 MR images corresponding to 110 patients where 2556 images are normal and 1373 images are with abnormalities. We take 10% of the data for validation and 10% for testing in the self-supervision task. Thereafter, a number of class balanced subsets of the previous training set are selected to learn the classification network according to Table 1.

Results and Analysis: Table 1 shows the average accuracy, macro F1, and AUROC over 5 runs with different random initializations. We evaluate the classification performance of all methods for three different training set sizes containing 961, 480, and 192 samples in each. Our method outperforms the rest in all metrics and training set sizes with a higher margin for lower training set sizes showing the robustness of our method. Other methods have similar performances

Table 1. Results of the compared methods on LGG dataset

Training data	Method	Accuracy	Macro F1	AUROC
192	Reconstruction	0.682 ± 0.0218	0.676 ± 0.0237	0.763 ± 0.0362
	Context Restoration	0.69 ± 0.0226	0.686 ± 0.0237	0.768 ± 0.037
	Context Prediction	0.732 ± 0.0063	0.731 ± 0.0063	0.81 ± 0.0179
	Intelligent-Masking	$\mathbf{0.764 \pm 0.0168}$	$\mathbf{0.761 \pm 0.0176}$	$\mathbf{0.819 \pm 0.0103}$
480	Reconstruction	0.744 ± 0.0241	0.743 ± 0.0246	0.812 ± 0.0322
	Context Restoration	0.75 ± 0.0059	0.749 ± 0.006	0.818 ± 0.0279
	Context Prediction	0.754 ± 0.0113	0.753 ± 0.0109	0.832 ± 0.0205
	Intelligent-Masking	$\mathbf{0.782 \pm 0.0075}$	$\mathbf{0.78 \pm 0.0073}$	$\mathbf{0.855 \pm 0.0165}$
961	Reconstruction	0.764 ± 0.0254	0.764 ± 0.0256	0.822 ± 0.0291
	Context Restoration	0.768 ± 0.0145	0.767 ± 0.0146	0.826 ± 0.0269
	Context Prediction	0.762 ± 0.0104	0.762 ± 0.0099	0.842 ± 0.0208
	Intelligent-Masking	$\mathbf{0.775 \pm 0.013}$	$\mathbf{0.774 \pm 0.0126}$	$\mathbf{0.855 \pm 0.0197}$

in large training sets; however, they perform differently for lower training sizes, from which, Context Prediction obtains better results than others.

3.2 Breast Cancer Detection

Detection of breast cancer and its type is our second experiment. Breast Ultra-Sound Images (BUSI) dataset [1] includes 2D ultrasound images collected from women between ages 25 and 75 years old in 2018. This dataset provides 780 single-channel grayscale ultrasound images categorized into 133 normal, and 647 cancer cases. We split the data into 50%, 25%, and 25% for training, validation, and test, respectively in the self-supervision task. Subsequently, a number of class balanced subsets of the previous training set are selected to learn the classification network according to Table 2.

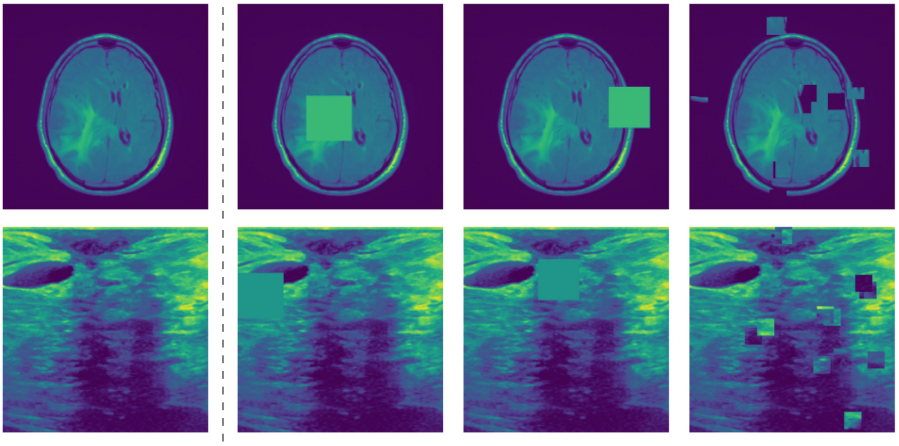
Results and Analysis: Table 2 shows the average accuracy, macro F1, and AUROC over 5 runs with different random initializations. We compare the performance of the methods on three different training sizes. However, unlike the dataset in Sect. 3.1, there are much fewer samples in each setup in this dataset. The results show that Intelligent-Masking performs better particularly with smaller training annotations. We observe that Context Restoration shows lower results compared to Context Prediction which might be due to the fact that our classification tasks need the pre-trained network focus on the local structures like tumor regions rather than the global information. Meanwhile, the loss of the Context Restoration is defined globally over the entire image unlike the Context Prediction and Intelligent-Masking that define the loss over the masked regions.

3.3 Qualitative Analysis

In Fig. 2, we present qualitative examples of different masking strategies. It is observed that, unlike context prediction and restoration, our method tends to

Table 2. Results of the compared methods on BUSI dataset

Training data	Method	Accuracy	Macro F1	AUROC
13	Reconstruction	0.576 ± 0.063	0.521 ± 0.1089	0.616 ± 0.0179
	Context Restoration	0.675 ± 0.0067	0.658 ± 0.008	0.668 ± 0.0127
	Context Prediction	0.752 ± 0.0344	0.744 ± 0.0425	0.801 ± 0.003
	Intelligent-Masking	$\mathbf{0.776 \pm 0.0106}$	$\mathbf{0.775 \pm 0.0103}$	$\mathbf{0.857 \pm 0.0079}$
66	Reconstruction	0.585 ± 0.0645	0.579 ± 0.0675	0.667 ± 0.0973
	Context Restoration	0.716 ± 0.0183	0.709 ± 0.0187	0.773 ± 0.0095
	Context Prediction	0.779 ± 0.0221	0.777 ± 0.0213	0.828 ± 0.0116
	Intelligent-Masking	$\mathbf{0.794 \pm 0.0221}$	$\mathbf{0.79 \pm 0.024}$	$\mathbf{0.862 \pm 0.0058}$
129	Reconstruction	0.639 ± 0.0267	0.638 ± 0.0272	0.702 ± 0.0098
	Context Restoration	0.71 ± 0.0403	0.707 ± 0.0366	0.788 ± 0.0067
	Context Prediction	0.776 ± 0.0236	0.774 ± 0.0249	0.827 ± 0.0062
	Intelligent-Masking	$\mathbf{0.8 \pm 0.0309}$	$\mathbf{0.798 \pm 0.0332}$	$\mathbf{0.865 \pm 0.005}$



(a) Original (b) Intelligent-Masking (c) Context Prediction (d) Context Restoration

Fig. 2. Qualitative examples of compared method's strategies for masking

propose targeted masks like the tumor regions or regions with abnormalities and avoids masking less helpful regions. It should be noted that Intelligent-Masking does not necessarily mask the tumor regions but considers all areas of interest that results in better feature learning. Examples of other masking samples are provided in supplementary materials. Furthermore, in medical images, unlike natural scenes, the structures are very local with imbalanced information throughout an image. Therefore, random masking strategies as shown in Fig. 2 operate ineffectively by masking non-informative regions.

4 Discussion and Conclusion

This paper proposes an intelligent masking strategy to train rich feature representation instead of random masking. To this aim, we leverage a reinforcement learning-based agent which learns to intelligently mask input images through deep Q-learning. To evaluate our model, we use the trained encoder for the classification task on two medical image datasets for the diagnosis of lower-grade glioma and breast cancer. Experimental results show that Intelligent-Masking has a consistent enhancement in terms of accuracy, macro F1, and AUROC compared to the SOTA methods. We believe that this work opens a path for enhancing self-supervision strategies like rotation and patch swapping and making them more context-aware.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)
2. Bai, W., et al.: Self-supervised learning for cardiac MR image segmentation by anatomical position prediction. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 541–549. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_60
3. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. Adv. Neural Inform. Process. Syst. **19** (2006)
4. Buda, M., Saha, A., Mazurowski, M.A.: Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Comput. Biol. Med. **109**, 218–225 (2019)
5. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration. Med. image Anal. **58**, 101539 (2019)
6. Da, K.: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1422–1430 (2015)
8. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
9. Hester, T., et al.: Deep q-learning from demonstrations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
10. Jamaludin, A., Kadir, T., Zisserman, A.: Self-supervised learning for spinal MRIs. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 294–302. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_34
11. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9664–9674 (2021)
12. Liao, R., et al.: An artificial agent for robust image registration. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
13. Liao, X., et al.: Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9394–9402 (2020)

14. Mazurowski, M.A., Clark, K., Czarnek, N.M., Shamsesfandabadi, P., Peters, K.B., Saha, A.: Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the cancer genome atlas data. J. Neuro-oncol. **133**(1), 27–35 (2017)
15. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
16. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)
17. Tajbakhsh, N., et al.: Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1251–1255. IEEE (2019)
18. Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D., Kainz, B.: Multiple landmark detection using multi-agent reinforcement learning. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 262–270. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_29
19. Watkins, C.J., Dayan, P.: Q-learning. Mach. Learn. **8**(3), 279–292 (1992)
20. Watkins, C.J.C.H.: Learning from delayed rewards (1989)
21. Xiong, J., et al.: Edge-sensitive left ventricle segmentation using deep reinforcement learning. Sensors **21**(7), 2375 (2021)
22. Zhang, M., Xu, J., Abaci Turk, E., Grant, P.E., Golland, P., Adalsteinsson, E.: Enhanced detection of fetal pose in 3D MRI by deep reinforcement learning with physical structure priors on anatomy. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 396–405. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_38
23. Zhang, P., Wang, F., Zheng, Y.: Self supervised deep representation learning for fine-grained body part recognition. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 578–582. IEEE (2017)
24. Zhou, S.K., Le, H.N., Luu, K., Nguyen, H.V., Ayache, N.: Deep reinforcement learning in medical imaging: a literature review. Med. Image Anal. **73**, 102193 (2021)
25. Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., Zheng, Y.: Self-supervised feature learning for 3d medical images by playing a rubik's cube. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 420–428. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_46

A New Lightweight Architecture and a Class Imbalance Aware Loss Function for Multi-label Classification of Intracranial Hemorrhages

Prabhat Lankireddy, Chitimireddy Sindhura, and Subrahmanyam Gorthi(✉)

Department of Electrical Engineering, Indian Institute of Technology Tirupati, Tirupati, Andhra Pradesh, India
prabhat.lankireddy@gmail.com, sindhurareddy234@gmail.com, s.gorthi@iittp.ac.in

Abstract. Deep learning algorithms have proven effective in solving many medical imaging tasks in recent years. The design of lightweight neural networks is gaining importance in the medical imaging community as not many hospitals and clinics are equipped with high computational resources to deploy large deep learning algorithms. Also, medical imaging data often comes with high class imbalance and thus there is a high necessity to develop deep learning models that can address this issue. With this motivation, a resource-efficient deep learning model called Lightweight-Fully Convolutional Network (LightFCN) is developed which can be deployed in clinical settings with limited computational resources. Label Distribution Aware Margin loss (LDAM) is used in the context of medical imaging for the first time for multi-label classification with class imbalance. The proposed model has a smaller memory footprint, a smaller number of parameters, lesser inference time and fewer Floating Point Operations (FLOPS) when compared to state-of-the-art models, without compromising on performance and can be deployed in clinical settings with limited computational resources. The model and the performance of the loss function are evaluated on the task of Intracranial Hemorrhage (ICH) classification on CT scans, and the model was deployed on a Raspberry Pi 4B (8 GB), on which inference times were compared. It is found that the proposed model significantly reduced the number of model parameters by a factor of 26, and reduced the inference time by a factor of 3, when compared to the popular lightweight network MobileNetV2.

Keywords: Deep learning · Class imbalance · Lightweight · Multi-label · Image classification · Intracranial hemorrhage

1 Introduction

Deep learning models perform very well on medical imaging tasks like classification and segmentation, but they are often over-parameterized, and lightweight

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 397–405, 2022.
https://doi.org/10.1007/978-3-031-21014-3_41

neural network design is all about looking for a network that can do the same task while consuming significantly less time and resources. Lightweight neural network architectures are now being given plenty of importance due to their multiple advantages, such as reduction of training time, inference time, lesser memory footprint and lesser number of computations (Floating Point Operations, or FLOPS). This is gaining importance in the medical imaging community as many hospitals and clinics are not equipped with high computational resources to deploy large deep learning algorithms, nor can they be replaced by new machinery owing to financial constraints.

Lightweight Neural Network design has been given a lot of importance in Computer Vision literature. A few popular architectures include MobileNet [7], which replaces regular convolutions with depthwise separable convolutions to reduce the number of FLOPS and inference time, and Shufflenet [18], that uses pointwise group convolutions and channel shuffling to reduce the number of FLOPS. The research focus in the medical imaging community is on resource efficient architectures, for various classification and segmentation tasks [12,16].

Class imbalance is often prevalent in medical imaging data because classification or segmentation tasks often involve identifying a lesion or disease, and the collection of huge data with rarer lesions or diseases is impractical. This causes the data to be skewed towards the "negative" samples i.e. samples without a lesion, and when deep learning models are trained on such data, they tend to learn to classify a test sample as negative more often. One easy way to address skewed data problems is to simply remove the excess data belonging to majority classes in an effort to have approximately equal number of data samples for all classes, but that is impractical in medical imaging tasks owing to the shortage of data in many cases, as less data makes it harder to train deep learning models that are extremely data hungry. This brings the need to develop methods to address the problem of class imbalance, in order to improve the performance of deep learning models for medical imaging tasks.

Lots of work has been done in addressing class imbalance when training machine learning algorithms, but the same cannot be said for deep learning tasks. A review paper on deep learning with class imbalance [8] broadly categorised the methods that address class imbalance into three different types, which are re-sampling methods, re-weighting methods and loss function based methods. Re-sampling involves over-sampling (training with the positive-labeled data more often than negative-labeled data) or under-sampling (skipping training on excess negative-labeled data), while re-weighting involves giving a higher weight to the losses for rare class data than losses for data of the frequently occurring classes, as done in the class balanced loss [4]. There are also loss-function based methods in which a loss function is designed to address class imbalance. Few examples for this are the focal loss [10], which gives a higher weight to the samples that are classified improperly by the model, and the Label-Distribution-Aware Margin Loss (LDAM) [3] which gives a higher classification margin to the rarer class samples than the dominant class samples.

While class imbalance has been discussed in the context of multi-class classification and segmentation tasks, little to no attention has been paid to multi-label

classification tasks. In multi-label classification, there are multiple classes and more than one of them can be positive for sample data. One example of multi-label classification tasks is the identification of different types of Intracranial Hemorrhages (ICH) in brain CT scans, where more than one type of injuries can be identified at once. This approach is more time and resource efficient than using many binary-classification models to identify each kind of injury.

In this paper, we propose a lightweight architecture called Light-Fully Convoluted Neural Network (Light-FCN) for the multi-label image classification task of ICH identification, along with ways to address class imbalance in the training data to improve performance of the model.

The rest of the paper is organized as follows. Section 2 discusses the architecture of the proposed model and the loss function. The Sect. 3 describes the dataset and the implementation details, followed by the ablation study and results. The conclusions are presented in Sect. 4.

2 Architecture and Loss Function

2.1 Lightweight Architecture Design

We propose a simple and lightweight architecture, *Light-FCN*, for the detection of ICH in CT scans. This architecture is inspired from the Simple Fully Convolutional Network (SFCN), proposed by Peng et al. [14], which has been found to work well in the task of Brain Age Prediction using CT scans. SFCN is developed based on the VGGNet [11] and the fully convolutional network (FCN) [17]. The success of SFCN in the task of Brain Age prediction suggests that the network is capable of extracting useful features from CT scans, which was the reason we decided to use SFCN as the baseline architecture for our task. The architecture of the proposed model is shown in the Fig. 1.

To reduce the complexity while maintaining the performance, two modifications are made to the baseline model. Firstly, the convolution layers have been replaced with a combination of depth-wise separable convolutional layers and pointwise 1×1 convolutional layers, an implementation inspired from the MobileNet architecture [7]. This has significantly reduced the total number of parameters. Secondly, inverted residual connections along with linear bottlenecks are introduced, which were adopted from MobileNetV2 [15]. The residual connections enabled easy training and improved performance without the need for deeper networks.

The Light-FCN network has 5 expansion blocks, in each of which the number of channels are increased using pointwise convolutions (called *expansion*), followed by depthwise convolutions. After that, another pointwise convolution operation is used to shrink the number of channels (called *projection*), and a residual connection is added to this output. Using expansions and projections as mentioned above significantly reduces the number of floating point operations, with a negligible reduction in performance, as the information in a feature map with large number of features can be represented using a lower dimensional subspace [15]. We use a constant bottleneck width of 32 channels in LightFCN.

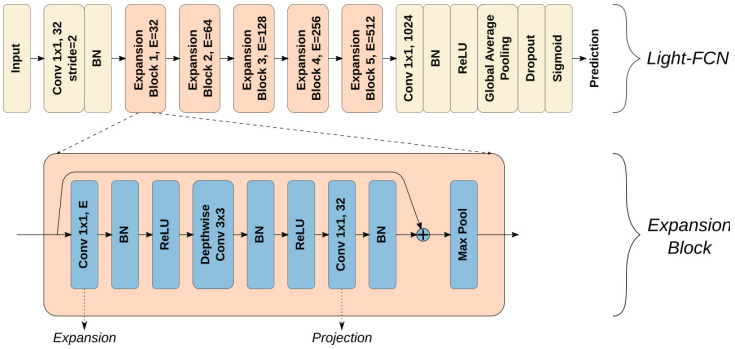


Fig. 1. Illustration of the proposed architecture Light-FCN and the Expansion block. "Conv 1x1, E" denotes a pointwise convolution layer with E output channels.

2.2 Loss Function

For multi-label classification, one can assume that all the outputs are independent of each other and devise a loss function accordingly. The sigmoid activation function at the output layer is suitable here, as it turns outputs into probabilities independent of other outputs. Another important factor to address when choosing a loss function, is the class imbalance in the training data. Certain loss functions are designed to minimize the bias that is introduced in the model due to the class imbalance present in the training data.

One such loss function is the LDAM loss function [3]. The LDAM loss introduces the concept of margin $\Delta_i = \frac{C}{n_i^{1/4}}$ (where C is a constant and n_i is the class frequency i.e. number of datapoints in the training data that belong to the i-th class.) and adds the margin to the model output if the samples corresponds to a negative ground truth (or subtracts the margin from the model if the samples correspond to a positive ground truth). To address class imbalance, a higher value of margin is used for classes that aren't well represented, leading to higher loss output. These modified loss outputs can be used along with any activation function. In this work, the sigmoid activation along with the binary crossentropy loss is chosen as we require the probabilities to be independent of each other. Then the loss is given as follows:

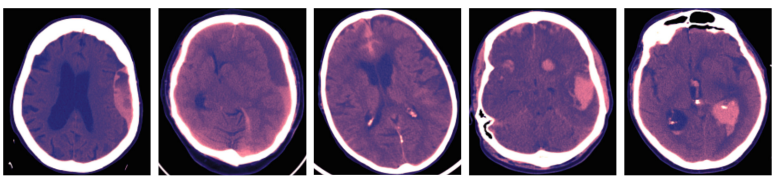


Fig. 2. Examples of preprocessed images for each type of injury

$$L(p,q) = -\frac{1}{k}\sum_{i}\left[p_i \log\left(\frac{1}{1+e^{-(z_i-\Delta_i)}}\right) + (1-p_i)\log\left(\frac{1}{1+e^{-(z_i+\Delta_i)}}\right)\right] \tag{1}$$

where p_i is the ground truth and z_i is the output logit for the i-th class.

3 Experiments and Results

3.1 Dataset and Preprocessing

The dataset used for this work is RSNA Intracranial Hemorrhage dataset [5]. Each CT slice was annotated with 5 binary labels indicating the presence of the following types of Intercranial Hemorrhage: 'Epidural Hemorrhage (EDH)', 'Intraparenchymal Hemorrhage (IPH)', 'Intraventricular Hemorrhage (IVH)', 'Subdural Hemorrhage (SDH)', and 'Subarachnoid Hemorrhage (SAH)'. The dataset has 21,744 CT scans in total, out of which only 7,652 had hemorrhages. As the injuries are observed in a small number of slices of a scan, only the scans that had injuries were included, in an effort to minimize class imbalance. The dataset has been split into 6122, 765 and 765 scans for training, validation and testing datasets respectively. The patient-wise and slice-wise distribution of data into train, validation and test datasets is shown in Table 1. All the 2D slices are down sampled to 256 × 256 and processed with intensity windowing method to construct a three channel RGB-like image. The three kinds of Hounsfield Units (HU) windows were considered for pre-processing as recommended in [2], each focusing on a different type of tissue: Brain window [Window Center (WC): 40, Window Width (WW): 80], Subdural window [WC: 80, WW: 200] and Soft Tissue window [WC: 40, WW: 380]. Few examples of each type of injury and the image obtained after pre-processing is shown in the Fig. 2.

Table 1. Patient-wise distribution of hemorrhages

	Training	Validation	Test
Patients	6,122	765	765
Slices	2,67,279	25,783	24,636
Without hemorrhages	1,78,052	16,211	15,502
With hemorrhages	89,227	9,572	9,134
Epidural	2,709	266	170
Intraparenchymal	29,860	3,301	2,957
Intraventricular	21,586	2,408	2,211
Subarachnoid	29,299	3,407	2,969
Subdural	38,864	4,166	4,136

3.2 Implementation Details

The proposed network has been implemented using Tensorflow [1] and has been trained on the RSNA dataset with a mini-batch size of 64. The network has been trained on a machine powered by 4 GeForce RTX 2080 Ti GPUs with 11GB memory each. No data augmentation was used during training. The Adam optimizer with an initial learning rate of 0.01 has been used to train the network. To prevent overfitting, a dropout layer with a probability of 0.1 is introduced to the final fully connected layer of the network. The values for mini-batch size, initial learning rate and dropout are obtained by carrying out hyperparameter optimization using the hyperband algorithm introduced by [9]. The KerasTuner library [13] has been used to automate the process of Hyperparameter optimization. The models are compressed using Tensorflow Lite [6], which is a framework that automates model compression for Tensorflow models and generated compressed models that can be deployed in edge devices like mobile phones or microcontrollers. The metrics that were used to measure task-related performance are Accuracy, Sensitivity, Specificity and Area under the ROC curve (AUC). To quantitatively measure computational efficiency, we obtain and compare the number of model parameters, memory footprint, inference time and the number of floating point operations per second (FLOPS). For all the latter, a lower number implies superior efficiency.

3.3 Ablation Study: From SFCN to LightFCN

Using the SFCN as the baseline model, a number of architectural modifications have been introduced to finally obtain the LightFCN model. The list of intermediate models along with the architectural modification is shown below:

1. *SFCN:* The Simple Fully Convoluted Network with no modifications.
2. *SFCN_depthwise:* All Convolution operations are replaced by Depthwise Separable Convolutions, as used in [7]. In addition to that, the first convolution layer is modified to have a stride of 2.
3. *LightFCN:* In addition to Depthwise Separable Convolutions, Linear Bottlenecks and Inverted residuals are added to the model as in [15].

All the models have been trained for 5 epochs, and the results are shown in Table 2. We can see that the architectural modifications hasn't degraded the performance by a lot, and yet managed to bring down the size by a significant amount (From 12.3 MB to 1.7 MB).

3.4 Comparing Performance with Other Models

The proposed method is compared with the state-of-the-art lightweight architectures: MobileNetV2 [15] and SFCN [14]. The models were all trained with same training data, and tested on same test data.

The quantitative comparison of the models using the aforementioned metrics is presented in Table 3. Our model achieved performance similar to the SFCN

Table 2. Comparison of the proposed Light-FCN model with the state-of-the-art methods.

Model	Accuracy	AUC	Specificity	Sensitivity	# Params	Size
SFCN	89.59	74.72	93.88	**55.55**	1 M	12.3 MB
SFCN_depthwise	89.85	69.68	94.38	44.97	136.1K	2.65 MB
Light-FCN	**92.09**	**74.76**	**97.99**	51.52	**86.8 K**	**1.7 MB**

Table 3. Performance comparison of the proposed model Light-FCN with the state-of-the-art architectures

MobileNetV2				
ICH	Accuracy	AUC	Sensitivity	Specificity
Epidural	96.66	62.10	27.06	97.14
Intraparenchymal	91.39	65.27	30.92	99.62
Intraventricular	92.36	89.22	85.40	93.05
Subarachnoid	88.56	69.92	45.36	94.47
Subdural	87.70	73.93	53.22	94.64
SFCN				
Epidural	98.37	56.24	13.53	98.96
Intraparenchymal	94.30	80.87	63.22	98.53
Intraventricular	95.39	77.75	56.26	99.24
Subarachnoid	91.34	69.49	40.71	98.26
Subdural	88.96	70.10	41.72	98.47
LightFCN				
Epidural	99.07	59.22	18.82	99.62
Intraparenchymal	90.63	76.28	57.42	95.14
Intraventricular	94.30	80.58	63.87	97.29
Subarachnoid	88.44	71.80	49.88	93.72
Subdural	84.23	67.55	42.47	92.64

model while reducing the number of parameters by a factor of 10. The sensitivity percentage for epidural hemorrhage is low for all models as a result of minimum number of samples in the training set. The highest sensitivity for SFCN and Light-FCN is obtained for intraventricular hemorrhage.

Furthermore, the models were created using the Tensorflow framework, and were compressed using the Tensorflow Lite framework [6] into the tflite format. Tensorflow Lite compresses a trained model through the means of quantization, clustering and pruning of model parameters. The resulting compressed models were deployed on a Raspberry Pi 4B, an accessible, low-cost, feature-rich embedded system, and their inference time was recorded. The sizes of the considered compressed models, their inference time, and the FLOPS required for each model are listed in Table 4. The experiments were performed on the proposed model with different loss functions and are listed in Table 5.

Table 4. Comparison of the compressed model size for the proposed model Light-FCN and different architectures along with their inference time (per slice) on Raspberry Pi

Model	Compressed model size	Inference time	#Params	FLOPS
MobileNetV2	8.46 MB	190.51 ms	2.3M	615M
SFCN	3.8 MB	174.79 ms	1M	2.07G
Light-FCN	**312 KB**	**61.15 ms**	**86.8K**	**189M**

Table 5. Evaluating performance of loss functions on the test dataset after training Light-FCN on the training dataset for 6 epochs.

Loss function	Accuracy	AUC	Sensitivity	Specificity
BCE	**93.49**	90.89	51.22	98.54
Focal	91.38	86.95	29.84	**98.87**
LDAM	92.44	**91.22**	**65.34**	94.62

3.5 Comparing Loss Functions

Using the LightFCN as the baseline model, the performance of multi-label versions of the binary crossentropy loss, focal loss and LDAM loss are compared. The multi-label binary crossentropy loss doesn't address class imbalance, but the multi-label focal loss and multi-label LDAM loss try to address class imbalance. In this study, three different LightFCN models are trained using each of the losses mentioned above and the performance metrics are observed.

It was observed that the LDAM loss had the highest sensitivity and AUC on the test dataset. Moreover, the LDAM loss seems to provide a good trade-off between sensitivity and specificity, thereby increasing the true positive rate in the presence of class imbalance, as expected. Interestingly, the binary crossentropy Loss gave better experimental results than the Focal Loss in accuracy, sensitivity and AUC. However, the LDAM loss seems to perform the best overall.

4 Conclusions

Our studies reveal that the Light-FCN architecture is capable of performing well in the task of ICH identification while being significantly more computationally efficient than other state-of-the-art architectures. The multi-label LDAM loss is observed to strike a better trade-off between sensitivity and specificity and increase the true positive rate, thereby solving the class imbalance problem to an extent. Achieving performance better than the state-of-the-art methods while maintaining computational efficiency remains the subject of future work.

Acknowledgements. This work is supported by the start-up research grant given by the Science and Engineering Research Board (SERB), India.

References

1. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous systems (2015). https://www.tensorflow.org/
2. Burduja, M., Ionescu, R.T., Verga, N.: Accurate and efficient intracranial hemorrhage detection and subtype classification in 3D CT scans with convolutional and long short-term memory neural networks. Sensors **20**(19), 5611 (2020)
3. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems 32 (2019)
4. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9268–9277 (2019)
5. Flanders, A.E., et al.: Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol. Artif. Intell. **2**(3), e190211 (2020)
6. Google: Tensorflow lite (2021). https://www.tensorflow.org/lite/
7. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
8. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. J. Big Data **6**(1), 1–54 (2019)
9. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A.: Hyperband: a novel bandit-based approach to hyperparameter optimization. J. Mach. Learn. Res. **18**(1), 6765–6816 (2017)
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007 (2017)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
12. Masud, M.: A light-weight convolutional neural network architecture for classification of COVID-19 chest X-ray images. Multimedia syst., 1–10 (2022)
13. O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al.: Kerastuner (2019). https://github.com/keras-team/keras-tuner
14. Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M.: Accurate brain age prediction with lightweight deep neural networks. Med. Image Anal. **68**, 101871 (2021)
15. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV 2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
16. Shuvo, M.B., Ahommed, R., Reza, S., Hashem, M.: CNL-UNet: a novel lightweight deep learning architecture for multimodal biomedical image segmentation with false output suppression. Biomed. Signal Process. Control **70**, 102959 (2021)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)



Spherical Transformer on Cortical Surfaces

Jiale Cheng[1,2], Xin Zhang[1], Fenqiang Zhao[2], Zhengwang Wu[2], Xinrui Yuan[2], John H. Gilmore[3], Li Wang[2], Weili Lin[2], and Gang Li[2(✉)]

[1] South China University of Technology, Guangzhou, China
[2] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
gang_li@med.unc.edu
[3] Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA

Abstract. Motivated by the recent great success of attention modeling in computer vision, it is highly desired to extend the Transformer architecture from the conventional Euclidean space to non-Euclidean spaces. Given the intrinsic spherical topology of brain cortical surfaces in neuroimaging, in this study, we propose a novel Spherical Transformer, an effective general-purpose backbone using the self-attention mechanism for analysis of cortical surface data represented by triangular meshes. By mapping the cortical surface onto a sphere and splitting it uniformly into overlapping spherical surface patches, we encode the long-range dependency within each patch by the self-attention operation and formulate the cross-patch feature transmission via overlapping regions. By limiting the self-attention computation to local patches, our proposed Spherical Transformer preserves detailed contextual information and enjoys great efficiency with linear computational complexity with respect to the patch size. Moreover, to better process longitudinal cortical surfaces, which are increasingly popular in neuroimaging studies, we unprecedentedly propose the spatiotemporal self-attention operation to jointly extract the spatial context and dynamic developmental patterns within a single layer, thus further enlarging the expressive power of the generated representation. To comprehensively evaluate the performance of our Spherical Transformer, we validate it on a surface-level prediction task and a vertex-level dense prediction task, respectively, i.e., the cognition prediction and cortical thickness map development prediction, which are important in early brain development mapping. Both applications demonstrate the competitive performance of our Spherical Transformer in comparison with the state-of-the-art methods.

Keywords: Transformer · Cortical surface · Cognition · Development prediction

1 Introduction

Convolutional Neural Network (CNN) based deep learning methods have dominated a wide range of tasks in computer vision for years and achieved tremendous success [8, 12, 13]. However, due to the inherent locality limitation presented in the convolutional operation, they lack the ability of modeling long-range dependencies explicitly. Recently, the pioneering work ViT [3] demonstrates the competitive performance of a

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 406–415, 2022.
https://doi.org/10.1007/978-3-031-21014-3_42

pure Transformer [25] architecture against CNNs in image classification. By computing pairwise interactions with a contextual aggregation mechanism across the whole image, this Transformer-based architecture outperforms its CNN counterparts with better capability to model long-range dependencies and parameter efficiency. This motivates an emerging interest in applying the powerful Transformer architecture in computer vision [7] and medical image analysis [1, 5, 10, 19, 24].

However, all these vision Transformers are designed for data that are represented in a grid-like Euclidean space, i.e., 2D/3D images. Therefore, they are inapplicable to complex anatomical shapes in medical imaging, where objects are typically represented with triangle meshes in a manifold space. For example, in neuroimaging analysis, the reconstructed brain cortical surfaces for precise representation and mapping of the cerebral cortex have highly convoluted shapes, but with an intrinsic spherical topology [16, 28]. Hence, in this paper, we seek to expand the applicability of Transformer such that it can serve as an effective general-purpose backbone for learning brain cortical surfaces, as it does for natural language processing and CNNs do for 2D/3D images.

Previous works on extending vision Transformers to image-based tasks in medical image analysis can be roughly divided into two lines according to their different policies on extracting low-level detailed information, which is an essential property in dense prediction tasks, such as segmentation and registration. The early studies [19, 24] adopted the first strategy in ViT [3], which embedded the non-overlapping image patches into tokens with a linear mapping directly. This strategy overlooks the detailed contextual information within the local patch and may result in an over-coarse feature map that cannot be effectively recovered by up-sampling to the original resolution. Moreover, the non-overlapping image patching scheme may introduce undesirable boundary effects in the outcome. To overcome these limitations, another type of methods [1, 5, 10] incorporated the rich local contextual information by employing a hybrid structure to build the Transformer on top of feature maps from a CNN backbone.

In this paper, we propose a novel Spherical Transformer, a pure Transformer architecture to capture the long-range dependency, while preserving the local contextual information on brain cortical surfaces. According to the prior spherical topology, an overlapping surface patching strategy is proposed. By conducting the within-patch self-attention mechanism, our Spherical Transformer effectively and efficiently extracts the fine-grained contextual information with linear computational complexity with respect to the patch size. Specifically, to build the cross-patch connection, we leverage the overlapping region, which inherently bridges neighboring patches and therefore enhances modeling power without any further processing, such as window shifting [18]. Together with the pooling and up-sampling layers in the spherical domain [29], a hierarchical representation is generated by gradually merging neighboring vertices in deeper Transformer layers and further integrated with advanced network structures, such as feature pyramid networks [17] and U-Net [22]. Moreover, we unprecedentedly propose a spatiotemporal self-attention (ST-SA) operation to fully leverage rich information in longitudinal surfaces. By computing the associative feature between pairwise vertices in patches from consecutive scans at the same position, the ST-SA operation can jointly extract the spatial context and growth patterns within a single layer, thus further enlarging the expressive power of the generated representation. To comprehensively validate the

performance of our Spherical Transformer, we test it on a surface-level prediction task, i.e., cognition prediction, and a vertex-level dense prediction task, i.e., cortical thickness map prediction. By comparing with the state-of-the-art algorithms, our proposed method demonstrates superior performance on both tasks.

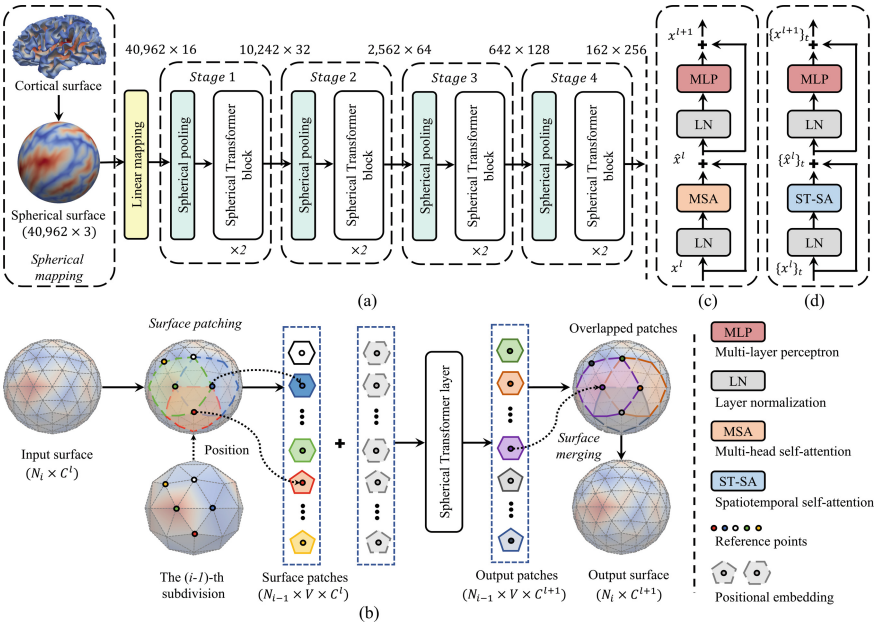


Fig. 1. Illustrations of (a) the backbone, (b) the Spherical Transformer block, which is consisted of the surface patching operation, Spherical Transformer layer, and the surface merging operation. Two types of Spherical Transformer layers are illustrated in (c) and (d), while (d) is designed for processing longitudinal data $\{x^l\}_t$. V denotes the number of vertices in the surface patch.

2 Method

An overview of our Spherical Transformer is presented in Fig. 1. It first splits an input cortical surface mapped on a sphere into overlapping patches by a surface patching module. The representation of each vertex in each patch is treated as a "token". Several Spherical Transformer layers with modified self-attention computation are applied on these tokens to compute the content-based feature aggregation and maintain the number of tokens. To produce a hierarchical representation, the pooling layer is adopted to produce high-level feature maps as the network gets deeper. In the following parts, we will introduce the surface patching strategy, our Spherical Transformer layer, spatiotemporal self-attention, and the overall architecture of our proposed Spherical Transformer.

2.1 Surface Patching Module

In vision Transformers, an image can be easily divided into patches for its grid-like Euclidean data structure. In contrast, a standard sphere for cortical surface representation is the subdivision of an icosahedron, which is generated starting from a regular icosahedron by iteratively adding a new vertex to the center of each edge in each triangle. Therefore, the number of vertices on the spherical surfaces are $N_1 = 12, N_{i+1} = 4N_i - 6$, $i = 1, 2, 3, \cdots$, where i presents the i-th subdivision. Existing methods typically adopted the non-overlapping splitting strategy, which will lead to undesirable boundary effects, and thus need further processing, e.g., window shifting [18]. Hence, in this work, we try to divide the cortical surface into overlapping regions instead. As shown in Fig. 1 (b), there are two types of surface patches with different shapes: (1) 12 patches in pentagon; and (2) the remainder in hexagon. In practice, we empirically partition a cortical surface with N_i vertices into N_{i-1} patches uniformly by selecting vertices on the $(i-1)$-th subdivision sphere as reference points and extracting their 2-hop neighborhoods to enlarge the receptive field within a single layer and sufficiently leverage the remarkable capability of the self-attention operation in formulating the long-range dependency. It can be observed that neighboring patches share around one third overlapping vertices. We compute the pair-wise associative features among vertices within the same patch by the self-attention operation to capture the low-level detailed context of cortical surfaces, while feature representations of overlapping vertices are calculated as the average of output patches. By partitioning cortical surfaces into overlapping patches, we introduce connections and feature transmission between neighboring patches which further smoothen the outcome in Spherical Transformer layers.

2.2 Spherical Transformer Layer

Given an input feature map $x^l \epsilon \mathbb{R}^{C^l \times V}$ of the l-th Transformer layer, where V denotes the number of vertices in the local patch, C denotes the length of feature representation of each vertex, the operation of a Transformer layer can be concluded as:

$$\hat{x}^l = MSA\left(LN\left(x^l\right)\right) + x^l, \tag{1}$$

$$x^{l+1} = MLP\left(LN\left(\hat{x}^l\right)\right) + \hat{x}^l, \tag{2}$$

where *MSA*, *MLP*, and *LN* denotes the multi-head self-attention operation, multi-layer perceptron, and the layer normalization operator, respectively.

To construct a more meaningful vertex representation as the input of our Spherical Transformer, we generate the vertex embedding by a linear mapping of the input feature vector at the beginning. Moreover, to encode the relative position of each vertex in the surface patch, we propose to learn specific positional embeddings $E_{pos} \in \mathbb{R}^{C^l \times V}$, which are added to vertex embeddings to retain positional information. As mentioned above, there are two types of patches on cortical surfaces with different shapes. Since there are only linear projections applied to surface patches in the Transformer layer, the number of vertices in each patch disentangles from the computation of the self-attention operation. Therefore, the Spherical Transformer layer can extract the detailed

contextual information of patches with both shapes without further modifications. On the other hand, we propose two different positional embeddings for patches in the pentagon and hexagon, respectively, for their different geometries as shown in Fig. 1 (b).

Computational Complexity. Attribute to the within-patch self-attention operation, the proposed Spherical Transformer layer enjoys great efficiency, compared to directly computing the associative feature among every vertex pair on the surface. Supposing we divide the spherical surface with N_i vertices into N_{i-1} patches, while the number of vertices in each patch is V, the computational complexity of a global Transformer layer (O_{global}) and the proposed Spherical Transformer layer (O_{patch}) are respectively:

$$O_{global} = 4N_i(C^l)^2 + 2(N_i)^2C^l \tag{3}$$

$$O_{patch} \approx 4N_i(C^l)^2 + 2N_{i-1}V^2C^l \tag{4}$$

where the former is quadratic to N_i, while the latter is linear when V is fixed. To be specific, when we set $N_i = 40,962$, $N_{i-1} = 10,242$ and $V = 19$ for patches in hexagon, the reduction ratio of the computational complexity is: $ratio = O_{patch}/O_{global} \approx 1/335$, with $C^l = 16$, which validates the efficiency of our network structure.

2.3 Spatiotemporal Self-Attention (ST-SA)

Aiming at designing a general-purpose backbone for cortical surfaces, processing the increasingly available longitudinal data is inevitable. To this end, we seek to extract the critical changing pattern within the Spherical Transformer layer rather than simply aggregate the temporal features before or after the Transformer backbone. An intuitive solution would be to expand the self-attention operation from 2D space (the spherical surface) into 3D (2D spatial + 1D temporal) space. We name it the spatiotemporal self-attention (ST-SA) operation, as illustrated in Fig. 1(d). It enables each vertex to directly interact with vertices in corresponding patches in consecutive scans, which greatly enlarges the model capability. Additionally, to indicate the order of different scans in the sequence, different positional embeddings are applied according to their temporal positions, which only introduces a minimal overhead to the model size.

2.4 Overall Architecture

To produce a hierarchical representation, as illustrated in Fig. 1(a), we stack multiple Spherical Transformer blocks with pooling layers and reduce the number of vertices on the surface from N_i to N_{i-1} as the network gets deeper, whereas we keep the patch size as the 2-ring neighborhood of N_{i-1} reference points. For surface-level prediction tasks, such as disease diagnosis and cognition prediction, we can simply employ this backbone and feed feature maps with the lowest resolution into a global average pooling layer and a fully connected layer. For the vertex-wise dense prediction tasks, such as cortical property map prediction and parcellation, the overall architecture is based on the U-Net [22] in an encoder-decoder paradigm. For processing longitudinal surface data, we will replace the MSA operation in the Spherical Transformer layer with our ST-SA

operation. As a result, the proposed architecture is very generic and can conveniently replace the backbone in existing methods for various tasks in brain cortical surface data analysis in the spherical domain.

3 Experiments

To validate the effectiveness of our proposed Spherical Transformer on cortical surfaces, we quantify its performance on two neurodevelopmentally important tasks, i.e., cognition prediction and cortical thickness map development prediction during infancy.

3.1 Dataset and Image Processing

We used the longitudinal T1w and T2w brain MR images from 110 healthy infants collected at their 0, 1, and 2 years of age [6]. All images were processed by an infant-tailored pipeline iBEAT V2.0 (http://www.ibeat.cloud) to generate cortical surfaces [14, 23, 26] and compute three morphological features of each surface vertex, including the cortical thickness, average convexity, and mean curvature. Then, these cortical surfaces were rigidly aligned and mapped onto the spherical space using FreeSurfer [4]. All spherical surfaces were aligned onto their age-matched templates in the 4D Infant Cortical Surface Atlas (https://www.nitrc.org/projects/infantsurfatlas/) [15, 27] and then resampled to have the same tessellation on the 6^{th} subdivision of icosahedron with 40,962 vertices.

To quantify the cognition development level of each participant, five Mullen cognitive scores [21] were estimated at their 2 years of age, i.e., Visual Receptive Scale (VRS), Fine Motor Scale (FMS), Receptive Language Scale (RLS), Expressive Language Scale (ELS), and Early Learning Composite (ELC), which are firmly correlated to the cortical morphological features mentioned above.

3.2 Model Configuration

These 110 longitudinal samples were randomly split into the training, validating, and testing sets with the proportion of 3:1:1. During the training phase, we used the Adam optimizer with L1 loss for the cortical thickness map development prediction. The learning rate was tuned cyclically within $[10^{-5}, 10^{-2}]$. The batch size was set to 1. For the cognition prediction task, we adopted the Adam optimizer with smooth L1 loss and tuned the learning rate cyclically within $[10^{-4}, 10^{-1}]$ with the batch size of 32.

Three different performance metrics were adopted to comprehensively evaluate the performance on cortical thickness map development prediction, i.e., mean absolute error (MAE), mean relative error (MRE), and the Pearson correlation coefficient (PCC), as in [20]. The cognition prediction task was evaluated by the root mean squared error (RMSE) on five different scales and their average.

3.3 Experimental Results

Two typical one-ring filter-based Spherical CNNs, i.e., SUNet [30] and UGSCNN [11] were selected as our baselines. We trained their models using the default network implementations provided in their source codes. Further, given that there is no prior work

applying Transformer on spherical cortical surfaces, we implemented another baseline, i.e., Hybrid-SUNet, which conducts the self-attention operation based on the output feature maps of SUNet, to comprehensively compare with the proposed method.

First, we evaluated our proposed Spherical Transformer method by using a single time point features (0 years of age) to predict the future outcome at 2 years of age on two different tasks. As shown in Tables 1 and 2, our Spherical Transformer achieved comparable results with other methods, including spherical CNNs and a hybrid architecture of CNN and Transformer in both tasks. In comparison between the Spherical Transformer and Hybrid-SUNet, the Spherical Transformer demonstrated superior performance, indicating the effectiveness and importance of capturing the long-range dependency for high-resolution feature maps with the attention-based mechanism.

Table 1. Comparison of different architectures for predicting cognitive scales at 2 years of age using cortical surface data at birth by root mean squared error (RMSE). * indicates statistically significant better results than other methods with p-value < 0.05.

Methods	VRS	FMS	RLS	ELS	ELC	Average
UGSCNN [11]	0.204	0.149	**0.153***	**0.148***	0.171	0.165 ± 0.028
SUNet [30]	0.186	0.139	0.163	0.161	0.167	0.163 ± 0.025
Hybrid-SUNet	0.186	0.138	0.163	0.163	0.167	0.164 ± 0.025
Proposed	**0.181***	**0.123***	0.159	0.166	**0.156***	$\mathbf{0.158 \pm 0.025}$*

Table 2. Comparison of different methods for predicting cortical thickness map at 2 years of age using cortical surface data at birth. * indicates statistically significant better results than other methods with p-value < 0.001.

Methods	MAE (mm)	MRE (%)	PCC
UGSCNN [11]	0.285 ± 0.023	15.15 ± 1.82	0.803 ± 0.029
SUNet [30]	0.246 ± 0.021	12.46 ± 1.30	0.853 ± 0.030
Hybrid-SUNet	0.248 ± 0.021	12.49 ± 1.33	0.851 ± 0.030
Proposed	$\mathbf{0.238 \pm 0.023}$*	$\mathbf{11.76 \pm 1.39}$*	$\mathbf{0.862 \pm 0.030}$*

Further, we applied our spatiotemporal self-attention in the Spherical Transformer to jointly model the spatial contextual information and temporal growth patterns. In Table 3, we tried to compare two model design strategies for longitudinal scans, i.e., aggregating the temporal dynamic features with a sequence model before the backbone, and jointly modeling the spatiotemporal information within the backbone. In this comparison, we selected our Spherical Transformer as the backbone for spatial feature extraction, and GRU [2], Transformer [25], and a 2-layer MLP as candidates for sequential modeling, which takes the longitudinal data of each vertex on the cortical surface as input and outputs its temporal developmental features. Longitudinal features from 0 and 1 year of

Table 3. Comparison of different temporal formulations by using longitudinal surface data at 0 and 1 year of age to predict the cortical thickness map (in terms of MAE, MRE and PCC) and cognitive scales (in terms of RMSE) at 2 years of age. * indicates statistically significant better results than other methods with p-value < 0.05.

Methods	MAE (mm)	MRE (%)	PCC	RMSE
Applied before the Spherical Transformer				
GRU [2]	0.207 ± 0.030	10.15 ± 1.61	0.889 ± 0.027	0.159 ± 0.025
MLP	0.209 ± 0.034	10.49 ± 1.90	0.881 ± 0.031	0.162 ± 0.024
Transformer [25]	0.212 ± 0.026	10.29 ± 1.31	0.889 ± 0.023	0.162 ± 0.023
Jointly extract spatiotemporal information				
ST-SA layer	**0.197 ± 0.039***	**9.60 ± 1.45***	**0.899 ± 0.025**	**0.150 ± 0.025***

age were utilized to predict the outcome at 2 years of age, such as the cognitive scales, and the cortical thickness map. As shown in Table 3, after introducing the longitudinal information, performances of both tasks are generally superior to that in Tables 1 and 2. An interesting observation is that if we separately describe the temporal pattern and the spatial context, in the temporal fusion part, Transformer [25] performs worse than the GRU and even the MLP. This may be caused by the inadequate time points we have, as the Transformer outperforms other sequential models in capturing the long-range dependency, which would be limited with only 2 time points. Therefore, our ST-SA operation substantially improved overall performance as it inherently enlarges the sequence length by combing the spatiotemporal receptive field. A visual comparison shown in Fig. 2 demonstrates the ST-SA layer with the Spherical Transformer predicts the cortical thickness map more precisely than other methods.

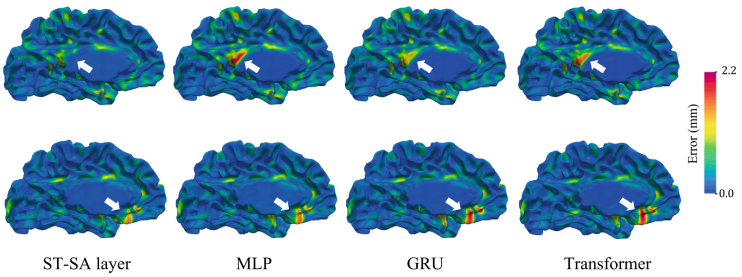


Fig. 2. Illustrations of the vertex-wise error maps of the cortical thickness development prediction of two randomly selected infants by different methods in terms of mean absolute error.

4 Conclusion

In this work, we proposed a novel Spherical Transformer, the first pure Transformer architecture, which can be applied to brain cortical surfaces with an intrinsic spherical

topology. By partitioning cortical surfaces into overlapping patches, it formulates the long-range dependency within the patch, bridges the neighboring patches with overlapping regions, and preserves the detailed contextual information, which is critical for dense prediction tasks. Since we compute the self-attention operation within the patch, our Spherical Transformer layer obtains linear computational complexity with respect to the patch size. Moreover, to extend the generalization ability of the Spherical Transformer to longitudinal data, we design the spatiotemporal self-attention operation to jointly capture the spatial contextual information and temporal development patterns. By testing our Spherical Transformer on two important yet challenging tasks, its superior performance has been validated against various spherical CNN methods and Transformer based methods. In the future, we will extensively test our Spherical Transformer on other cortical/subcortical surface tasks and make it publicly available.

Acknowledgements. This work was supported in part by the National Institutes of Health (NIH) under Grants MH116225, MH117943, and MH123202.

References

1. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
2. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Fischl, B.: Freesurfer. Neuroimage **62**(2), 774–781 (2012)
5. Gao, Y., Zhou, M., Metaxas, D.N.: Utnet: a hybrid transformer architecture for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 61–71. Springer (2021)
6. Girault, J.B., et al.: Cortical structure and cognition in infants and toddlers. Cereb. Cortex **30**(2), 786–800 (2020)
7. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al.: A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. (2022). https://doi.org/10.1109/TPAMI.2022.3152247
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)
10. Ji, Y., Zhang, R., Wang, H., Li, Z., Wu, L., Zhang, S., et al.: Multi-compound transformer for accurate biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 326–336. Springer (2021)
11. Jiang, C., Huang, J., Kashinath, K., Marcus, P., Niessner, M., et al.: Spherical CNNs on unstructured grids. arXiv preprint arXiv:1901.02039 (2019)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)

13. LeCun, Y., et al.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **1**(4), 541–551 (1989)
14. Li, G., et al.: Measuring the dynamic longitudinal cortex development in infants by reconstruction of temporally consistent cortical surfaces. Neuroimage **90**, 266–279 (2014)
15. Li, G., Wang, L., Shi, F., Gilmore, J.H., Lin, W., Shen, D.: Construction of 4D high-definition cortical surface atlases of infants: methods and applications. Med. Image Anal. **25**(1), 22–36 (2015)
16. Li, G., et al.: Computational neuroanatomy of baby brains: a review. Neuroimage **185**, 906–925 (2019)
17. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
19. Matsoukas, C., Haslum, J.F., Söderberg, M., Smith, K.: Is it time to replace cnns with transformers for medical images? arXiv preprint arXiv:2108.09038 (2021)
20. Meng, Y., Li, G., Gao, Y., Lin, W., Shen, D.: Learning-based subject-specific estimation of dynamic maps of cortical morphology at missing time points in longitudinal infant studies. Hum. Brain Mapp. **37**(11), 4129–4147 (2016)
21. Mullen, E.M., et al.: Mullen scales of early learning. AGS (ed.). Circle Pines, MN: American Guidance Service Inc. (1995)
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
23. Sun, L., Zhang, D., Lian, C., Wang, L., Wu, Z., Shao, W., et al.: Topological correction of infant white matter surfaces using anatomically constrained convolutional neural network. Neuroimage. **198**, 114–124 (2019)
24. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 36–46. Springer (2021)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
26. Wang, L., Li, G., Shi, F., Cao, X., Lian, C., Nie, D., et al.: Volume-based analysis of 6-month-old infant brain MRI for autism biomarker identification and early diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 411–419. Springer (2018)
27. Wu, Z., Wang, L., Lin, W., Gilmore, J.H., Li, G., Shen, D.: Construction of 4D infant cortical surface atlases with sharp folding patterns via spherical patch-based group-wise sparse representation. Hum. Brain Mapp. **40**(13), 3860–3880 (2019)
28. Zhao, F., Wu, Z., Li, G.: Deep learning in cortical surface-based neuroimage analysis: a systematic review. Intell. Med. (2022). https://doi.org/10.1016/j.imed.2022.06.002
29. Zhao, F., et al.: Spherical deformable U-Net: application to cortical surface parcellation and development prediction. IEEE Trans. Med. Imaging **40**(4), 1217–1228 (2021)
30. Zhao, F., Xia, S., Wu, Z., Duan, D., Wang, L., Lin, W., et al.: Spherical U-Net on cortical surfaces: methods and applications. In: International Conference on Information Processing in Medical Imaging, pp. 855–866. Springer (2019)



Accurate Localization of Inner Ear Regions of Interests Using Deep Reinforcement Learning

Ana-Teodora Radutoiu[1(✉)], François Patou[2], Jan Margeta[3], Rasmus R. Paulsen[1], and Paula López Diez[1]

[1] DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark
s194324@student.dtu.dk
[2] Oticon Medical, Research and Technology group, Smørum, Denmark
[3] KardioMe, Research and Development, Nova Dubnica, Slovakia

Abstract. We propose a novel method for automatic ROI extraction. The method is implemented and tested for isolating the inner ear in full head CT scans. Extracting the ROI with high precision is in this case critical for surgical insertion of cochlear implants. Different parameters, such as CT equipment, image quality, anatomical variation, and the subject's head orientation during scanning make robust ROI extraction challenging. We propose to use state-of-the-art communicative multi-agent reinforcement learning to overcome these difficulties. We specify landmarks specifically designed to robustly extract orientation parameters such that all ROIs have the same orientation and include the relevant anatomy across the dataset. 140 full head CT scans were used to develop and test the ROI extraction pipeline. We report an average overall estimated error for landmark localization of 1.07 mm. Extracted ROI presented an intersection over union of 0.84 and a Dice similarity coefficient of 0.91.

Keywords: Region of interest · Deep reinforcement learning · Computed tomography · Inner ear · Landmarks · Orientation

1 Introduction

Automatic region of interest (ROI) detection in medical images is a challenging task as medical images generally present high variability between individuals, scanners, and image acquisition and postprocessing protocols. ROI extraction is a necessary step for almost all medical image analysis pipelines. It is also vital for subsequent image processing tasks that rely on input stability. Accurate ROI extraction can not only improve retrieval efficiency but can also help to classify more easily pathological signs within a reduced region especially when the anatomy is particularly challenging for either the clinicians or the processing software to interpret [13].

Cochlear implants (CI) are used to restore the hearing capacities of patients who suffer from severe to profound hearing loss. CIs are common for infants with

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 416–424, 2022.
https://doi.org/10.1007/978-3-031-21014-3_43

congenital deafness. Clinicians evaluate each patient using computed tomography (CT) images, the head orientation is important to assess each case and successfully obtain the relevant measurements for accurate surgical planning. Therefore we had developed an approach to automatically locate this region in clinical full head CT images.

Initially, ROI detection methods were based on bounding boxes from handcrafted features [3,11]. Nowadays, the most common technique is deep learning, where a majority of methods in the literature are designed for 2D medical images [2,13] and other approaches used 2D methods to locate ROIs in 3D [5,8,16]. These methods fail to use the third dimensionality of CT images and need a big amount of annotated data that faithfully represents the anatomical variability and struggle when the anatomy is abnormal. We chose to use a deep reinforcement learning (DRL) based approach that uses landmarks (easier and faster to annotate) to automatically extract the inner ear ROI from CT images. DRL has been successful showing outstanding performance for similar tasks of landmark localization in medical images [1,15]. Other DRL approaches include Navarro et. al. [10], this paper proposes single reinforcement learning agents for organ localization in the torso. They succeeded in effectively finding a ROI around desired organs with a relatively small amount of data. We chose to use a landmark-based approach which we consider will be more robust and could potentially provide more explainability or even help detect abnormalities from an early processing stage [6,13].

2 Data

This study uses 140 full head CT scans from the CQ500 dataset [4] which corresponds to 102 different patients. The dataset has been provided by the Centre for Advanced Research in Imaging, Neurosciences and Genomics (CARING), New Delhi, India [4]. The CT scans are taken from several radiology centers in New Delhi and are collected using various equipment models [4]. All used scans are resampled to have the isotropic voxel spacing 0.5 mm. The image dimensions vary significantly within our dataset. On average the dimensions are $475 \times 475 \times 323$, but the dimensions $x \in [400; 576]$, $y \in [400; 576]$ and $z \in [128; 730]$. All scans are manually labeled with the chosen landmarks using the software 3D Slicer [7]. All the annotations are made public and can be found in https://github.com/AnaTeodoraR/annotations.git.

Choosing relevant landmarks is necessary to characterize the inner ear orientation. The landmarks must be uniquely defined within their structure, so they are easily differentiated from other anatomical points nearby. Eleven landmarks are chosen in total, five assigned for each inner ear ROI and one common for both. The five landmarks for each ROI are the same anatomical points but located on their respective side of the CT scan. All landmarks are associated with a number; Numbers 1–5 are in connection to the right ROI, 6–10 for the left, and 11 is common. Two landmarks are within the inner ear; cochlear apex (nr. 1 and 6) and superior semi-circular canal peak (nr. 4 and 9). Additionally, two landmarks in the cochlea nerve; the midpoint below the base of the cochlea

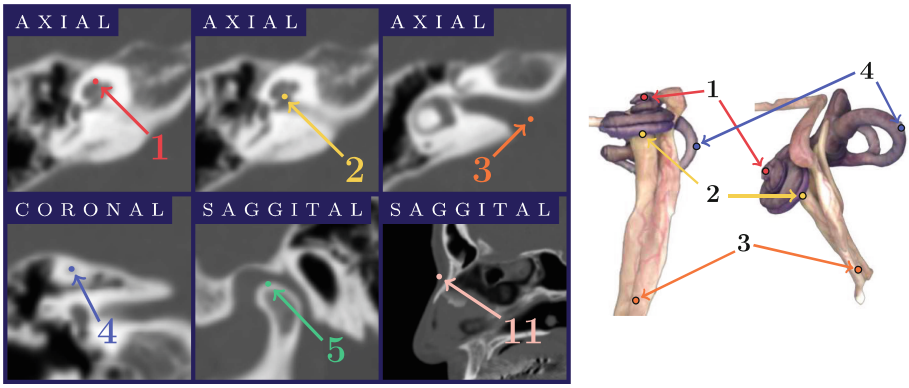


Fig. 1. Landmarks 1–5 and 11 are shown on CT scan nr. 6 and additionally landmarks 1–4 are shown on a 3D representation of an inner ear (edited from [14]). Landmarks 1–5 are for the right inner ear ROI, but the corresponding landmarks 6–10 are placed at the same anatomical points on the left side of the head.

(nr. 2 and 7) and of both the CN and FN further down (nr. 3 and 8). To aid in obtaining a similar orientation the remaining landmarks are; At the top of the condyle in the mandible where the temporomandibular joint starts (nr. 5 and 10) and lastly the nasion (nr. 11) (Fig. 1).

3 Methods

The strategy used for detecting the landmarks is DRL specifically the communicative multi-agent reinforcement learning (C-MARL) model proposed by Leroy et al. [9]. This model uses DRL with multiple agents that communicate and are based on a deep Q-learning network (DQN). The environment is defined as the entire 3D medical scan while an agent is moved voxel by voxel. A DQN is used to predict the optimal Q values given a certain state. As input the network takes states - a 3D patch centered around the agent - and outputs the Q values for each of the possible actions - 3D movement (up, down, left, right, forward, and backward) [9]. The architecture for two agents is shown in Fig. 2. All the agents share the weights of the convolutional layers (implicit communication) while each agent has its own fully connected (FC) layers only sharing the average output from each FC layer (explicit communication) [9]. The agents use multi-scale which enables them to have a spatially higher view of the image in its state. In our implementation, the agents used four scales including the isotropic voxel spacings 3, 2, 1, and 0.5 mm for the patch to represent the state. Initially, the agents will observe the states with the coarsest spacing, but when they start oscillating the spacing will decrease to the next, finer, resolution. The final model uses 11 agents, one per landmark, thus resulting in 11 FC layers and a discount rate of 0.8. The data is divided into training, validation, and test sets each being 112, 14, and 14 CT scans respectively (≈80%, ≈10%, and ≈10%).

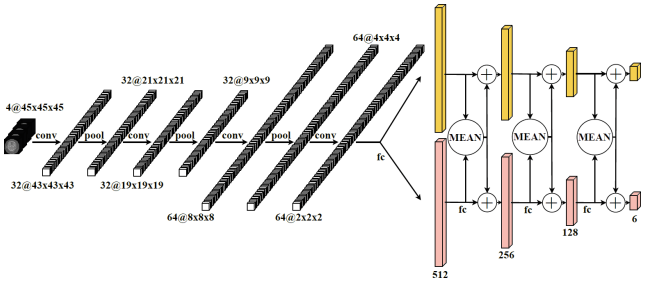


Fig. 2. Visualization of the CNN architecture, when using two agents. The fully connected layers for each agent are colored differently. Edited from [9].

The model was trained on a 12 GB GPU for five days. The training uses an ϵ-greedy strategy ($\epsilon \in [0.1; 1[$). The model which has the best performance on the validation set is further used for testing.

The aim of this study is to extract the two ROIs with similar orientation in a two-step approach. The first step aims to rotate the images to align all heads. The landmarks are used to define the relevant rotation angles that characterize the head orientation. The second step is a regular axis-aligned crop of the inner ear on the newly rotated image. To rotate a 3D medical image, three rotation angles denoted α, β, and γ are used, also known as yaw, roll, and pitch respectively. They describe the head movement in 3D as shown in Fig. 3.

The rotation angles α and β have the potential to be estimated based on the assumption that the corresponding left and right anatomical structures are symmetrical as is the case for the ear anatomy [12]. Technically, α and β are found by projecting the line between the same two anatomical points onto the axial plane and coronal plane respectively, and finding the deviation from the horizontal line, see Fig. 3a and 3b. Four independent estimations of these angles are found using the landmark pairs (1, 6), (2, 7), (3, 8), and (5, 10) and a median of these estimations is applied as the final rotation angle. Likewise, γ is found assuming landmark 5 or 10 and landmark 11 are at an angle of $\theta = 20°$ when projected on the sagittal plane as seen in Fig. 3c. Humans are anatomically different but we estimate 20° to be a good estimation of this anatomy. Finally, γ is defined as the angle which corrects the difference between θ and the estimated ones using the landmarks (Fig. 3c). Once more two different γ values can be estimated, so a median of these will be used as the final γ.

After obtaining the three angles for a single 3D medical scan, the image is rotated accordingly using three individual rotation transformations. The ROI is extracted on the rotated image as a 3D axis-aligned crop from a center point with a customized size. The center point is set to the middle point between two landmarks (landmarks 1 and 2 for the right, 6 and 7 for the left). The radii, r_x, r_y and r_z, of the 3D box, are found as the furthest distance to a set of landmark's respective x, y, and z coordinates (landmarks 1, 2, 3, and 4 for the right ROI,

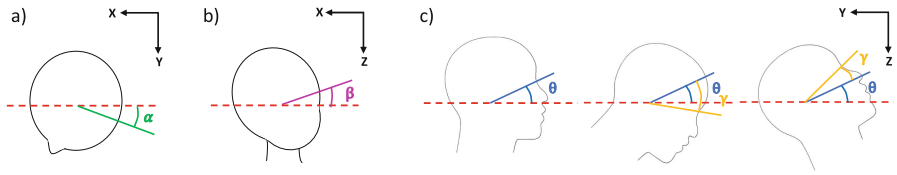


Fig. 3. Sketch of α, β, and γ where the rotation angles have been exaggerated. The angle $\theta = 20°$ describes the desired head position for γ. a) Axial view b) Coronal view c) Sagittal view and γ for the two different orientations.

while 6, 7, 8, and 9 for the left). Moreover, $r_x = r_y = \max(r_x, r_y)$, which results in the ROIs having quadratic axial slices.

A robustness analysis is made by applying a grid of different rotations to an image. The chosen image is the one where the sum of absolute calculated angles (from the annotations) is the smallest, hence the image closest to the reference pose. This image has the estimated angles; $\alpha = 0.19°$, $\beta = 1.08°$, $\gamma = 0.70°$, hence its sum is 1.97°. For our robustness analysis, we exhaustively sampled all 3D rotations on a uniform grid: $\alpha', \beta' \in \{-15, -10, -5, 0, 5, 10, 15\}$ and $\gamma' \in \{-20, -10, 0, 10, 20, 30\}$, where α', β', and γ' are the applied rotation along the third, second and first axis respectively. This results in $7 \cdot 7 \cdot 6 = 294$ manually rotated images. For all generated images their landmarks will be predicted using the C-MARL model. Furthermore, the angle calculation method is applied using the predicted landmarks and compared to the applied rotation. Consequently, the applied rotation will have the opposite sign of the predicted (if predicted correctly).

4 Results

Different C-MARL models have been trained and it was empirically found that using a discount rate of 0.8 and a single agent per landmark, performed best. Using multiple agents per landmark showed only to have a relevant influence on detecting landmark 11. The results of the final model on the test set can be seen in Table 1. Observing the table, the model performs particularly well at detecting landmarks within the inner ear and cochlear nerve (landmarks 1, 2, 3, 4, 6, 7, 8, and 9). For these eight landmarks, the mean distance error is below one millimeter. Depending on the landmark, the estimated errors vary between 0.79–2.11 mm.

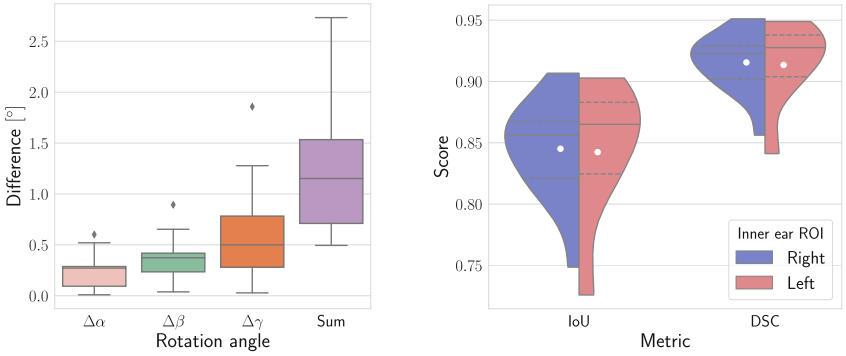
Figure 4a shows box-plots of the rotation angle differences across the test images. Here the rotation angle difference is between the predicted angles and the ones found from the landmark annotations. The maximum deviation of the predicted angle from the estimated (calculated from the annotated landmarks) for α and β is below 1° and the difference for γ is capped at 1.8°, with its upper quartile is below 1°. Figure 4a additionally illustrates a box-plot of the sum of the three differences in a test image which ranges from 0.36°–2.68°.

Now we evaluate the ROI extraction performance. Figure 4b illustrates the intersection over union (IoU) and Dice similarity coefficient (DSC) for both ROIs, additionally, we show the distribution between right and left ROIs. The IoU and DSC are found by comparison with an estimated ROI using the annotated landmarks. Observing Fig. 4b, a small difference between the right and left ROI's prediction exists. The scores for the right and left ROI appear unimodal having a single peak.

The results of the robustness analysis are shown in Fig. 5. Observing the two figures, each 3D point represents a manually rotated image whose three applied rotation angles (α', β', and γ') are its x, y, and z coordinates. The points are color-coded depending on the performance. Looking at Fig. 5a, many of the rotated images have an estimated error of 1–2 mm (purple color). A general tendency of more purple colors for smaller and negative γ' is observed and a decline in performance is seen as γ' increases. The lowest accuracy tends to be gathered at highly negative α', positive β', and positive γ' rotations. Similar patterns can be observed in the corresponding figure for the angle error (Fig. 5b). The error here is the sum of differences between the predicted rotation angles

Table 1. The estimated error (mean distance error, d_{Error}) and standard deviation (σ_{Error}) on the test set. The model predicts the landmarks three times on each test image and a median is used as the final prediction. The last row shows the percentage of detections below one millimeter.

Landmark	1	2	3	4	5	6	7	8	9	10	11	Overall
mean(d_{Error}) [mm]	0.84	0.79	0.87	0.87	1.63	0.83	0.85	0.82	0.99	2.11	1.24	1.07
σ_{Error} [mm]	0.38	0.42	0.64	0.44	1.14	0.28	0.39	0.56	0.31	1.04	0.64	0.75
<1 mm [%]	64.3	78.6	78.6	64.3	35.7	78.6	57.1	85.7	57.1	7.1	50	–



(a) Box-plots of angle differences.

(b) Violin-plots of ROI scores.

Fig. 4. Left: Box-plot of the angle differences on the test images (both the individual ones and their sum in an image). **Right:** Violin-plot of IoU and DSC on the test images. The line represents the median, the stripes the upper and lower quartile, and the white dot the mean.

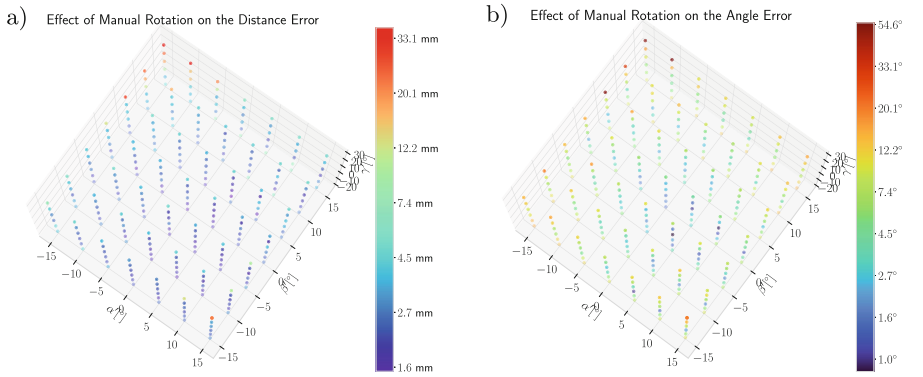


Fig. 5. Each point represents an image that has been manually rotated, the rotation angles used are illustrated as its x, y, and z coordinates. The colors illustrate the estimated error for a landmark (right) and the sum of the angle prediction errors (left). The coloring has a logarithmic scale. (Color figure online)

and the applied (with opposite sign). Only a few rotated images have an angle error below 2° (blue and black colors), while the general error tends to be around 4–10° (green, lime, and yellow colors).

5 Discussion

We found the landmark accuracy shown in this work sufficient for the ROI extraction task. Studying the performance of the individual landmarks, a few points can be made. The agents looking for landmarks 1–4 (and 6–9) are searching close by one another, and potentially benefiting more from the communication. Thus, a reason as to why the highest precision for landmark localization is achieved for these eight landmarks. Interestingly, landmark 11 is isolated from the rest, hence why the single agent has more trouble detecting it. Landmarks 5 and 10 emerge as the most difficult to locate. These landmarks mark the peak of the condyle where the jaw joint starts. The peak is a slightly curved surface, moreover, if a patient has a rotating head when taking the CT scan, other parts of the mandible become visible at the same axial slice. These matters make landmarks 5 and 10 difficult to place, both for the annotator and agents.

Multi-scale is of importance [1], especially if agents are looking for a small structure within a big image. It is important that the agents quickly get an idea of which region the landmark is in. The multi-scale enables the agents to view a larger section of the image and quickly narrow down which region their landmark is located. We implemented 4 levels of multi-scale that were sufficient for our task. The preciseness of the landmark localization influences the angle prediction. Moreover, γ is predicted using the landmarks with the lowest localization precision (landmarks 5, 10, and 11), why this rotation angle has the lowest prediction performance of the three (Fig. 4a).

Regarding the robustness analysis in Fig. 5, negative γ' corresponds to rotating the head downwards in the sagittal plane. This is a more common position of the head, why the model performs better at these rotations. Additionally, $\pm 10°$ with α' and β' further worsens the performance, however, it is also less common for a patient to have these rotations when taking a CT scan. Thus, the results of the robustness analysis reveal that rotations that mirror usual head positions perform best. Data augmentation could potentially help the model to be more robust with extreme head rotations.

Considering the ROI extraction, the overall predictions have on average an IoU score of 0.84 and DSC of 0.91. Since the IoU penalizes false positives and false negatives more than the DSC, it explains the lower scores observed for the IoU. Additionally, the regions are not all fixed to be the same size, so it might be the case that a predicted ROI is contained within an estimated (or oppositely). These cases result in a lower IoU and DSC. We compared our ROI extraction method with another state of the art method as seen in Table 2. Table 2 compares the highest average IoU (the liver) achieved by Navarro et. al. [10] with the left and right ROI results. Both inner ear ROIs have on average a higher IoU than the liver ROI. Our presented method is specifically designed to extract the inner ears with likewise orientation. On the other hand, Navarro et. al. [10] strives to achieve a general axis-aligned ROI detection framework (using DRL) for organs in the torso.

Table 2. Comparing the average IoU for the left and right ROI with the average IoU for the liver (best organ result) from Navarro et. al.

	Navarro et al. [10]	This model right ROI	This model left ROI
Avg IoU	0.80	0.836	0.835

6 Conclusion

This study successfully used a DRL framework for landmark detection in full head CT scans, and utilize it for ROI extraction of the inner ears. Landmarks were localized with an estimated error between 0.78–2.11 mm (on average 1.07 mm) within this difficult anatomical structure. The defined rotation angles gave the ROIs the desired orientation. Two ROIs were extracted from the detected landmarks with an overall average IoU of 0.84 and DSC of 0.91. The method outperforms other DRL approaches for ROI detection as is the proposed by Navarro et. al. [10]. Through this study, we explored the capability of implementing C-MARL for predicting fine structures in full head CT scans. This paves the way for analysis of inner ear ROIs for surgical use.

References

1. Alansary, A., et al.: Evaluating reinforcement learning agents for anatomical landmark detection. Med. Image Anal. **53**, 156–164 (2019). https://doi.org/10.1016/j.media.2019.02.007

2. Bi, L., Kim, J., Kumar, A., Fulham, M., Feng, D.: Stacked fully convolutional networks with multi-channel learning: application to medical image segmentation. Vis. Comput. **33**(6), 1061–1071 (2017)
3. Campadelli, P., Casiraghi, E., Esposito, A.: Liver segmentation from computed tomography scans: a survey and a new algorithm. Artif. Intell. Med. **45**(2–3), 185–196 (2009)
4. Chilamkurthy, S., et al.: Development and validation of deep learning algorithms for detection of critical findings in head CT scans (2018). https://doi.org/10.48550/ARXIV.1803.05854, dataset. http://headctstudy.qure.ai/dataset
5. De Vos, B.D., Wolterink, J.M., De Jong, P.A., Viergever, M.A., Išgum, I.: 2D image classification for 3D anatomy localization: employing deep convolutional neural networks. In: Medical imaging 2016: Image processing, vol. 9784, pp. 517–523. SPIE (2016)
6. Diez, P.L., et al.: Deep reinforcement learning for detection of abnormal anatomies. In: Proceedings of the Northern Lights Deep Learning Workshop, vol. 3 (2022)
7. Fedorov, A., et al.: 3D slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging **30**(9), 1323–1341 (2012). https://doi.org/10.1016/j.mri.2012.05.001
8. Hiraman, A., Viriri, S., Gwetu, M.: Efficient region of interest detection for liver segmentation using 3D CT scans. In: 2019 Conference on Information Communications Technology and Society (ICTAS), pp. 1–6 (2019). https://doi.org/10.1109/ICTAS.2019.8703625
9. Leroy, G., Rueckert, D., Alansary, A.: Communicative reinforcement learning agents for landmark detection in brain images. In: Kia, S.M., et al. (eds.) MLCN/RNO-AI -2020. LNCS, vol. 12449, pp. 177–186. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-66843-3_18. https://arxiv.org/abs/2008.08055
10. Navarro, F., Sekuboyina, A., Waldmannstetter, D., Peeken, J.C., Combs, S.E., Menze, B.H.: Deep reinforcement learning for organ localization in ct. In: Arbel, T., et al. (eds.) Proceedings of the Third Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 121, pp. 544–554. PMLR, 06–08 July 2020. https://proceedings.mlr.press/v121/navarro20a.html
11. Peng, J., Hu, P., Lu, F., Peng, Z., Kong, D., Zhang, H.: 3D liver segmentation using multiple region appearances and graph cuts. Med. Phys. **42**(12), 6840–6852 (2015)
12. Reda, F.A., McRackan, T.R., Labadie, R.F., Dawant, B.M., Noble, J.H.: Automatic segmentation of intra-cochlear anatomy in post-implantation CT of unilateral cochlear implant recipients. Med. Image Anal. **18**(3), 605–615 (2014). https://doi.org/10.1016/j.media.2014.02.001
13. Sudha, S., Jayanthi, K., Rajasekaran, C., Sunder, T.: Segmentation of RoI in medical images using CNN-a comparative study. In: TENCON 2019–2019 IEEE Region 10 Conference (TENCON), pp. 767–771. IEEE (2019)
14. Trier, P., Noe, K.: The visible ear simulator (2020). https://ves.alexandra.dk/
15. Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D., Kainz, B.: Multiple landmark detection using multi-agent reinforcement learning. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 262–270. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_29
16. Ying, W., Cunxi, C., Tong, J., Xinhe, X.: Segmentation of regions of interest in lung CT images based on 2-D Otsu optimized by genetic algorithm. In: 2009 Chinese Control and Decision Conference, pp. 5185–5189. IEEE (2009)

Shifted Windows Transformers for Medical Image Quality Assessment

Caner Özer[1(✉)], Arda Güler[2], Aysel Türkvatan Cansever[2], Deniz Alis[3], Ercan Karaarslan[3], and İlkay Öksüz[1,4]

[1] Istanbul Technical University, Istanbul, Turkey
ozerc@itu.edu.tr
[2] Mehmet Akif Ersoy Thoracic and Cardiovascular Surgery Training and Research Hospital, Istanbul, Turkey
[3] Acibadem University, Istanbul, Turkey
[4] King's College London, London, UK

Abstract. To maintain a standard in a medical imaging study, images should have necessary image quality for potential diagnostic use. Although CNN-based approaches are used to assess the image quality, their performance can still be improved in terms of accuracy. In this work, we approach this problem by using Swin Transformer, which improves the poor-quality image classification performance that causes the degradation in medical image quality. We test our approach on Foreign Object Classification problem on Chest X-Rays (Object-CXR) and Left Ventricular Outflow Tract Classification problem on Cardiac MRI with a four-chamber view (LVOT). While we obtain a classification accuracy of 87.1% and 95.48% on the Object-CXR and LVOT datasets, our experimental results suggest that the use of Swin Transformer improves the Object-CXR classification performance while obtaining a comparable performance for the LVOT dataset. To the best of our knowledge, our study is the first vision transformer application for medical image quality assessment.

Keywords: Visual transformers · Medical image quality assessment · Chest x-rays · Cardiac MRI

1 Introduction

It is the responsibility of technicians to ensure that the acquired medical scan should be of good quality so that the physician, who asked for the medical scan, will be able to diagnose accurately. However, this condition may not be satisfied due to the medical image quality problems such as patient breathing or arrhythmia in Cardiac Magnetic Resonance Imaging (MRI), ringing artifact presence in Computed Tomography (CT), and low dose or foreign object appearance on the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_44.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 425–435, 2022.
https://doi.org/10.1007/978-3-031-21014-3_44

Chest X-Rays (CXR). These image quality issues would also constitute a critical problem for automated processing of medical scans within a clinical decision support system consisting of big data.

Maintaining a high standard in image quality is imperative also for a machine learning-guided medical image analysis downstream tasks. More specifically, existence of poor quality images may lead a machine learning methodology to learn spurious set of features which are not clinically useful. Hence, identifying, eliminating or correcting the samples with poor image quality is necessary for downstream tasks such as segmentation and disease diagnosis. The reasons for an image that should be eliminated from the pipeline include but are not limited to artefacts in MR Imaging [11], foreign object existence [15], or undesirable region appearance [12] within the scans. For this purpose, several datasets on a variety of modalities have been developed to overcome such problems. Ma et al. [11] has developed a brain MRI dataset in which the patient data has been labelled according to the level of motion artefact. The authors have also developed a 4-layer convolutional neural network model to classify the samples according to their motion degradation level. JFHealthcare [8] proposed a MIDL 2020 challenge which comes up with a dataset that aims to localize the foreign objects within chest X-Rays.

Swin-Transformers are becoming more common in medical image analysis, especially in the field of diagnosis. For instance, Zhang et al. [18] uses a cascaded mechanism for chest CT COVID-19 diagnosis which uses a U-Net to extract lungs as a foreground and background segmentation mechanism, and Swin-Transformer uses these segmented lungs to find out whether the patient has COVID-19 or not. Chen et al. [2] combines a statistical significance-based slice sampling scheme and Swin-Transformers to propose a framework that can detect COVID-19 from chest CT scans by using a single slice. In addition to diagnosis, Swin-Transformers are becoming more common to be used for other problems, such as brain tumor segmentation [1,5], and cardiac CT reconstruction [13]. However, despite being promising for these problems in medical imaging, Swin-Transformers' have yet to be studied for medical image quality assessment tasks.

In this work, we propose a SwinTransformer-based [10] approach for medical image quality analysis where our contributions are as follows:

- We train a variety of SwinTransformer models in order to classify X-Ray images with foreign objects and Cardiac MRI images with a 4-chamber view with Left Ventricular Outflow Tract (LVOT).
- We compare our approach with deep neural network (DNN) based approaches, such as ResNet [6] and EfficientNet [14].
- To our best knowledge, this is the first paper to perform medical image quality assessment through SwinTransformers.

2 Methodology

2.1 Pre-processing

In order to train our model, we apply a number of preprocessing procedures prior to feeding the input images to the network. In this regard, instead of man-

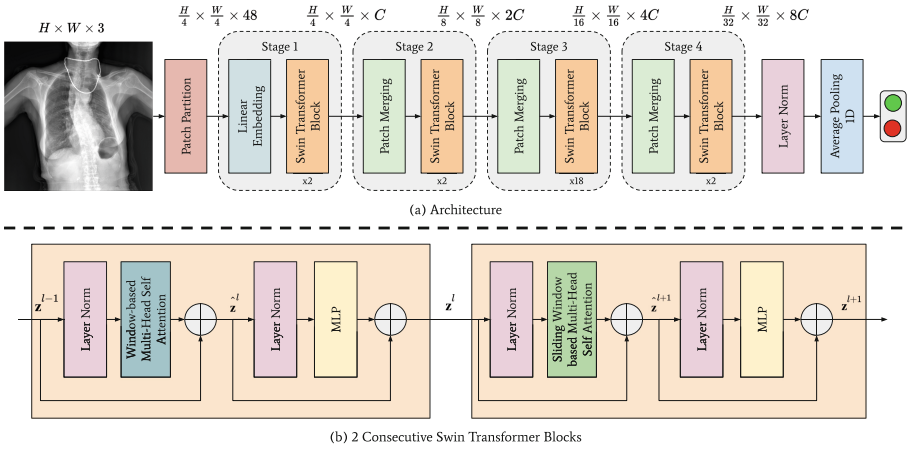


Fig. 1. Swin Transformer pipeline for medical image quality analysis.

ually defining the data augmentations and their corresponding hyperparameters, we use RandAugment [3] which uses a previously found automated data augmentation strategy to train a DNN. After that, we resize these images to a pre-determined resolution of $H \times W$. Finally, we normalize the images by using ImageNet statistics that they become ready-to-use in a pre-trained model.

2.2 Swin Transformer

Following the pre-processing procedure, we used the Swin Transformer model that is being demonstrated in Fig. 1. Given an input image, Swin Transformer first divides the image into patches using a Patch Partition layer that generates 48-dimensional representations for each of the 4×4 patches on the image. After that, as shown in Stage 1, the Linear Embedding layer transforms 48-d representations into C dimensions.

The second component of Stage 1 is the Swin Transformer Blocks which has been inherited from the general Visual Transformer (ViT) [4] architecture. Despite having the same high-level block structure as the one in ViT, multiheaded self-attention (MSA) blocks are replaced with an alternate ordering of window-based MSA (W-MSA) and shifted window-based MSA (SW-MSA), sequentially. In W-MSA, the feature map is divided into windows and MSA computes the output on each of them. However, this brings a cost of missing global connectivity, since this was not an issue for MSA in ViT that operates globally. For this reason, cyclic shifting operation is used to overcome this problem with SW-MSA that, the feature map is circularly shifted on the height and width axes by the half-length of the patch size. Then, W-MSA is executed on the resulting feature map, and a reverse circular shift operation is used to obtain a reverted feature map without any shifts.

In all of Stages 2, 3, and 4, we have a Patch Merging layer and a varying number of Swin Transformer blocks. In the Patch Merging layer, each of the pixels located on the 2×2 patches on the feature map are concatenated across the channel axis, i.e., $\frac{H}{2} \times \frac{W}{2} \times C$ to $\frac{H}{4} \times \frac{W}{4} \times 4C$. Then, after applying the layer normalization layer, a linear layer is used to reduce the dimensionality to half, i.e., to $\frac{H}{4} \times \frac{W}{4} \times 2C$. The output is then transferred to the Swin Transformer blocks for further processing. Lastly, after using layer normalization followed by an average pooling layer to calculate the features of the Swin Transformer model, the output is used for classification in a single Linear layer. This model classifies if there is anything present in the input image that can affect the image quality adversely. The hyperparameter settings can be found in Table 1. Stochastic depth refers to the probability of disregarding the transformation that is applied on a sample within a batch during training, while being processed within a Swin Transformer block, embedding dims C is the dimensionality of the patch-wise embeddings, depths and MSA heads correspond to the number of repeated Swin Transformer blocks and head count of the MSA layer for each stage, respectively.

Table 1. Hyperparameter settings for different Swin Transformer models.

Model name	Stochastic depth	Embedding dims (C)	Depths	MSA heads
Swin-Tiny	0.2	96	2, 2, 6, 2	3, 6, 12, 24
Swin-Small	0.3	96	2, 2, 18, 2	3, 6, 12, 24
Swin-Base	0.2	128	2, 2, 18, 2	4, 8, 16, 32

2.3 Training the Model

We adopt the official code of Swin Transformers[1] in which we performed some modifications to run it on PyTorch 1.10.2. We also used TorchVision 0.11.3 to inherit the pre-trained models for EfficientNet and ResNet pre-trained architectures. We apply the learning rate warm-up scheme and learning rate scheduling and leverage color jittering, stochastic depth (only for Swin Transformer) [7], Random Erasing [19] MixUp [17], CutMix [16] and RandAugment [3] data augmentation methods for promoting model generalization. For the foreign object classification model, we set the batch size to 16, gradient accumulation steps to 8 and window size to 8. Instead for the LVOT classification model, we set the batch size to 64, gradient accumulation steps to 2 and window size to 7, due to lower input image resolution comparing chest X-Rays. We employ an AdamW [9] optimizer with weight-decay of 10^{-8} to fine-tune both of these models by using an ImageNet pretrained model, as we set the initial learning rate to 0.00006 and use a linear learning-rate scheduler. The fine-tuning procedure takes 60 epochs with 5 epochs of a warm-up period. We also trained the models from scratch for 300 epochs that, the learning rate is set to 0.0005 and the objective function is Cross Entropy. You can access to our repository by using the following link: https://github.com/canerozer/qct.

[1] https://github.com/microsoft/Swin-Transformer.

3 Experimental Results

3.1 Datasets

We demonstrate our results on the Object-CXR [8], a challenge dataset where the aim is to classify the Chest X-Ray scans with foreign objects and to localize the foreign objects, and the LVOT classification dataset, which is constructed by selecting a subset of the Cardiac MR (CMR) scans with 4-chamber view that were acquired in the Mehmet Akif Ersoy Thoracic and Cardiovascular Surgery Training and Research Hospital between the years 2016 and 2019. Object-CXR dataset contains 8,000 samples for training and 1,000 samples for each of the validation and testing splits. There is an equal number of positive and negative samples for each split which does and does not contain foreign objects. We only use the class label, but not the localization ground truth for this dataset during training. We set the fixed input size to $H = W = 1024$ since the images can reach a spatial resolution of 4048×4932. On the other hand, LVOT classification dataset contains 272 positive and 278 negative samples that are used for training, 34 positive and 35 negative samples for validation, and 35 positive and 35 negative samples for testing. These images have relatively way less spatial resolution than the ones in Object-CXR, hence, we set $H = W = 224$. In addition, since our pipeline is capable of processing 2D images, we slice the 3D CMR patient scans into 2D images, which is also helpful in terms of increasing the number of samples for each split by a factor of 25.

3.2 Left Ventricular Outflow Tract Classification Results

First, we start by comparing the performances of different Swin Transformer models that are named as Swin-Tiny, Swin-Small, and Swin-Base in Table 2. We additionally measured the differences between using a pretrained model and training from scratch. We notice that, the accuracy is the highest for the Swin-Tiny model with an accuracy of 95.59%. Nevertheless, Swin-Small's AUC (Area Under Curve) score is performing the best among these three models with a score of 0.980. For this reason, we decided to include both of the Swin-Small and Swin-Tiny models to compare them with the other baselines. Meanwhile, switching from pre-trained model to training from scratch significantly decreases the accuracy of at least around 12.35% and AUC score of 0.059. We relate this to the data-hungry regime of Swin Transformers, as they need abundant data samples to capture the inductive bias during training. Hence, it becomes necessary to use the pretrained models or to increase the number of samples to the order of tens of thousands.

Table 2. Comparison among Swin Transformer models with different capacity for the LVOT classification task.

Model	Pretrained?	Acc	AUC	# of params	GFLOPs	Throughput (img/sec)
Swin-Tiny	Yes	**95.59**	0.969	27,520,892	4.5	515
	No	72.93	0.757			
Swin-Small	Yes	95.48	**0.980**	48,838,796	8.7	313
	No	83.13	0.921			
Swin-Base	Yes	93.86	0.976	86,745,274	15.4	222
	No	67.01	0.738			

Table 3. Our results on the LVOT dataset indicate that the performance is comparable to the other DNN-based methods.

Model	Acc	AUC	# of params	GFLOPs	Throughput
EfficientNet-B0	93.74	0.979	4,010,110	0.4	2274
EfficientNet-B1	93.68	0.971	6,515,746	0.6	1565
EfficientNet-B2	94.61	0.971	7,703,812	0.7	1480
ResNet-34	95.19	0.974	21,285,698	3.7	1963
ResNet-50	95.77	0.979	23,512,130	4.1	1005
ResNet-101	95.59	0.969	42,504,258	7.8	587
ResNet-152	**96.00**	**0.988**	58,147,906	11.6	409
Swin-Tiny	95.59	0.969	27,520,892	4.5	515
Swin-Small	95.48	0.980	48,838,796	8.7	313

Then, we compare our approach with other deep learning baselines as ResNet [6] and EfficientNet [14] as shown in Table 3. We see that, despite the differences in the number of parameters and GFLOPs, the performance varies insignificantly ($p > 0.05$) with outstanding accuracy and AUC metrics. We relate this to the triviality of the LVOT classification task, where the goal is to check if fifth chamber exists on the Cardiac MRI scan - generally located at the centre. As a result, it becomes possible to obtain testing accuracy of over 95% and AUC scores exceeding 0.969 for all ResNet architectures and Swin-Small without overfitting. Still, these models are ready to be deployed for a real-world scenario for alerting LVOT regions on 4-chamber Cardiac MRIs almost instantaneously.

3.3 Foreign Object Classification Results

The results show Swin Transformer's strength on a higher dimensional input, namely, on the Chest X-Ray data for foreign object classification. The model complexity is highly correlated with the model performance that, the best performance is obtained by the Swin-Base model with 87.1% testing accuracy and 0.922 testing AUC score. The closest performance to the best model is the ResNet-152 model, which has 1.5× less number of parameters and 25% less floating

point operations, at a cost of 2% less testing accuracy and 0.013 less AUC score. Despite Swin-Base is working slightly slower than real-time due to its huge input size, its significant performance gains ($p < 0.005$) makes it a promising approach (Table 4).

Table 4. Results for the Object-CXR dataset.

Model	Acc	AUC	# of params	GFLOPs	Throughput
EfficientNet-B0	82.1	0.873	4,010,110	8.3	119
EfficientNet-B1	83.0	0.882	6,515,746	12.3	83
EfficientNet-B2	83.2	0.890	7,703,812	14.2	78
ResNet-34	82.9	0.895	21,285,698	91.8	94
ResNet-50	82.8	0.897	23,512,130	102.7	42
ResNet-101	84.1	0.906	42,504,258	195.8	26
ResNet-152	85.1	0.909	58,147,906	241.5	22
Swin-Base	**87.1**	**0.922**	86,766,330	324.6	11

3.4 Qualitative Analysis

Prior to utilizing our model in a clinical study, we further examine our model under extreme conditions. Firstly, we start with the Object-CXR classification task as some cases are shown in Fig. 2. Regardless of the orientation and contrast factors, our model can correctly classify the X-Ray images which does and does not contain foreign objects as demonstrated in Fig. 2a and 2b, respectively. What is interesting especially for the rightmost image of Fig. 2b is that, even though the foreign objects may not be visible to human eye without altering the contrast, our model can still successfully classify the image as it contains foreign objects. Foreign objects are marked with a blue bounding box in Fig. 2b.

We also perform a qualitative analysis on the classification results as some of the edge cases are shown in Fig. 3 for the LVOT classification task. In Fig. 3a, where all 3 images are labelled as good quality images without LVOT appearance, the model can correctly classify even when the images are subject to motion artefacts and local noise upto some level. Also, as shown in Fig. 3b, our model can correctly classify the images containing LVOT regions under low contrast. We show the regions with LVOT with a blue bounding box in Fig. 3b.

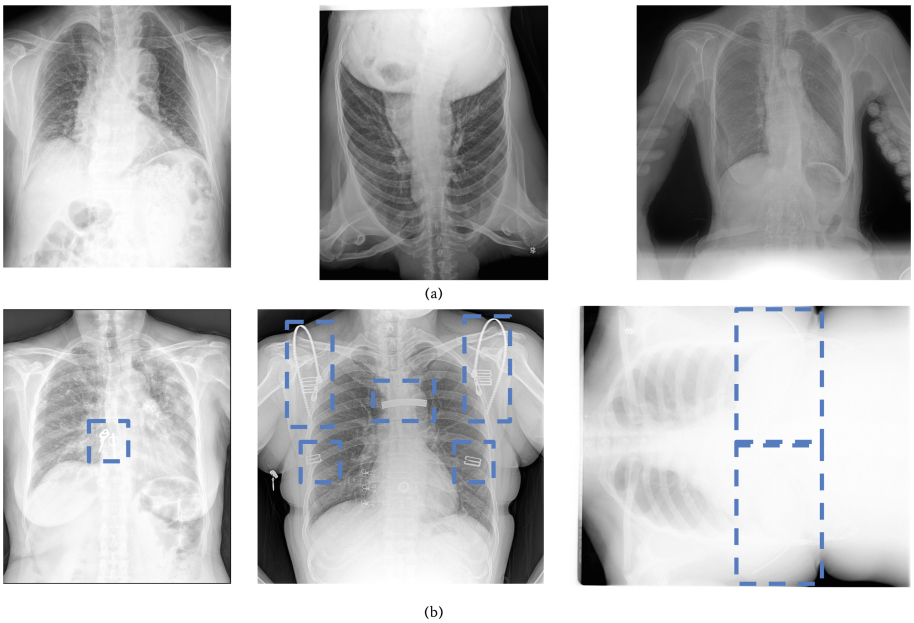


Fig. 2. Some true challenging samples due to their medical image quality factors, from Object-CXR dataset.

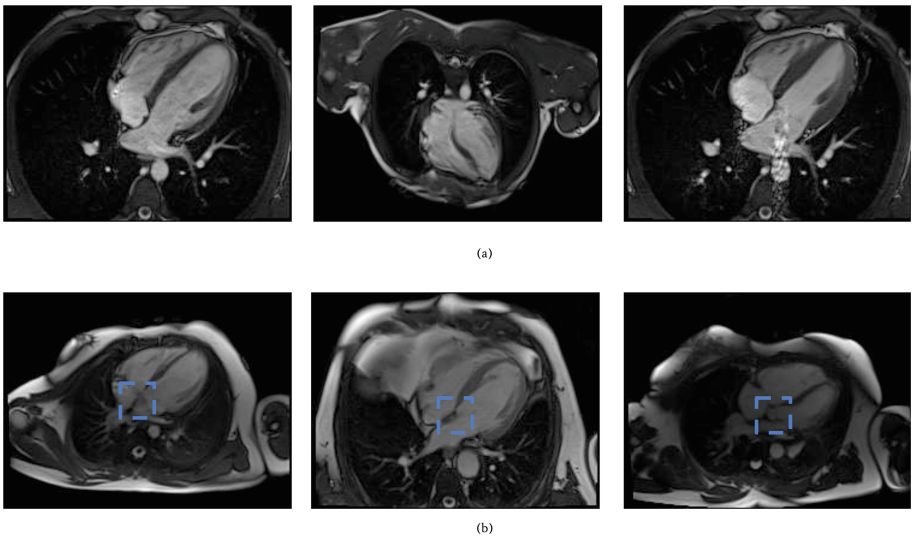


Fig. 3. Some true challenging samples due to their medical image quality factors, from LVOT dataset.

4 Discussion and Conclusion

In this work, we propose to use Swin Transformers for medical image quality analysis, where we validated our approach on two datasets working on two different modalities. We outperform a variety of ResNet and EfficientNet baselines on the Object-CXR dataset by obtaining a testing accuracy of 87.1% while obtaining a comparable performance for the LVOT classification objective. Comparing Swin Transformer to the other baselines and performing a qualitative analysis demonstrate its potential for utilizing it in a clinical setting.

Although the problem setting has been restricted to foreign object classification in Chest X-Rays and LVOT classification for Cardiac MRI's, it is possible to generalize the approach for other medical image quality problems such as motion artefact grade classification. In addition, the interpretability aspect of the Swin Transformers is a critical avenue of research for the safe transition of an automated medical image analysis pipeline.

Acknowledgments. This paper has been produced benefiting from the 2232 International Fellowship for Outstanding Researchers Program of TUBITAK (Project No: 118C353). However, the entire responsibility of the publication/paper belongs to the owner of the paper. The financial support received from TUBITAK does not mean that the content of the publication is approved in a scientific sense by TUBITAK.

References

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-Unet: Unet-like pure transformer for medical image segmentation (2021)
2. Chen, G.L., Hsu, C.C., Wu, M.H.: Adaptive distribution learning with statistical hypothesis testing for COVID-19 CT scan classification. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 471–479 (2021). https://doi.org/10.1109/ICCVW54120.2021.00057
3. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: RandAugment: practical automated data augmentation with a reduced search space. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 18613–18624. Curran Associates, Inc. (2020). https://proceedings.neurips.cc/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021. OpenReview.net (2021). https://openreview.net/forum?id=YicbFdNTTy
5. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. CoRR abs/2201.01266 (2022). https://arxiv.org/abs/2201.01266

6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 770–778. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.90
7. Huang, G., Sun, Yu., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 646–661. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_39
8. JFHealthcare: Automatic detection of foreign objects on chest x-rays. https://academictorrents.com/details/fdc91f11d7010f7259a05403fc9d00079a09f5d5 (2020). Accessed 27 Nov 2020
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015. Conference Track Proceedings (2015). http://arxiv.org/abs/1412.6980
10. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022, October 2021
11. Ma, J.J., et al.: Diagnostic image quality assessment and classification in medical imaging: opportunities and challenges. In: 17th IEEE International Symposium on Biomedical Imaging, ISBI 2020, Iowa City, IA, USA, 3–7 April 2020, pp. 337–340. IEEE (2020). https://doi.org/10.1109/ISBI45749.2020.9098735
12. Oksuz, I., et al.: Automatic left ventricular outflow tract classification for accurate cardiac MR planning. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 462–465 (2018). https://doi.org/10.1109/ISBI.2018.8363616
13. Pan, J., Wu, W., Gao, Z., Zhang, H.: MIST-net: multi-domain integrative Swin transformer network for sparse-view CT reconstruction. CoRR abs/2111.14831 (2021). https://arxiv.org/abs/2111.14831
14. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, 9–15 June 2019. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (2019). http://proceedings.mlr.press/v97/tan19a.html
15. Xue, Z., et al.: Foreign object detection in chest x-rays. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 956–961 (2015). https://doi.org/10.1109/BIBM.2015.7359812
16. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: CutMix: regularization strategy to train strong classifiers with localizable features. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), 27 October–2 November 2019, pp. 6022–6031. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00612
17. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018, Conference Track Proceedings. OpenReview.net (2018). https://openreview.net/forum?id=r1Ddp1-Rb

18. Zhang, L., Wen, Y.: A transformer-based framework for automatic COVID19 diagnosis in chest CTS. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 513–518 (2021). https://doi.org/10.1109/ICCVW54120.2021.00063
19. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, 7–12 February 2020, pp. 13001–13008. AAAI Press (2020). https://aaai.org/ojs/index.php/AAAI/article/view/7000



Multi-scale Multi-structure Siamese Network (MMSNet) for Primary Open-Angle Glaucoma Prediction

Mingquan Lin[1], Lei Liu[2], Mae Gorden[2], Michael Kass[2], Sarah Van Tassel[1(✉)], Fei Wang[1(✉)], and Yifan Peng[1(✉)]

[1] Weill Cornell Medicine, New York, NY, USA
{sjh2006,few2001,yip4002}@med.cornell.edu
[2] Washington University School of Medicine in St. Louis, St. Louis, MO, USA

Abstract. Primary open-angle glaucoma (POAG) is one of the leading causes of irreversible blindness in the United States and worldwide. POAG prediction before onset plays an important role in early treatment. Although deep learning methods have been proposed to predict POAG, these methods mainly focus on current status prediction. In addition, all these methods used a single image as input. On the other hand, glaucoma specialists determine a glaucomatous eye by comparing the follow-up optic nerve image with the baseline along with supplementary clinical data. To simulate this process, we proposed a Multi-scale Multi-structure Siamese Network (MMSNet) to predict future POAG event from fundus photographs. The MMSNet consists of two side-outputs for deep supervision and 2D blocks to utilize two-dimensional features to assist classification. The MMSNet network was trained and evaluated on a large dataset: 37,339 fundus photographs from 1,636 Ocular Hypertension Treatment Study (OHTS) participants. Extensive experiments show that MMSNet outperforms the state-of-the-art on two "POAG prediction before onset" tasks. Our AUC are 0.9312 and 0.9507, which are 0.2204 and 0.1490 higher than the state-of-the-art, respectively. In addition, an ablation study is performed to check the contribution of different components. These results highlight the potential of deep learning to assist and enhance the prediction of future POAG event. The proposed network will be publicly available on https://github.com/bionlplab/MMSNet.

Keywords: Deep learning · Primary open-angle glaucoma (POAG) · Fundus photographs · Siamese network

1 Introduction

Primary open-angle glaucoma (POAG) is one of the leading causes of blindness worldwide [1]. In the United States, POAG is the most common form of glaucoma and is the leading cause of blindness among African-Americans [22] and Hispanics [11]. POAG can be asymptomatic until very advanced stages.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 436–445, 2022.
https://doi.org/10.1007/978-3-031-21014-3_45

Fortunately, most blindness caused by POAG can be avoided by early identification and treatment [23]. Therefore, early prediction of eyes that will develop POAG plays an important role in patient monitoring and medical and surgical treatments [5,21].

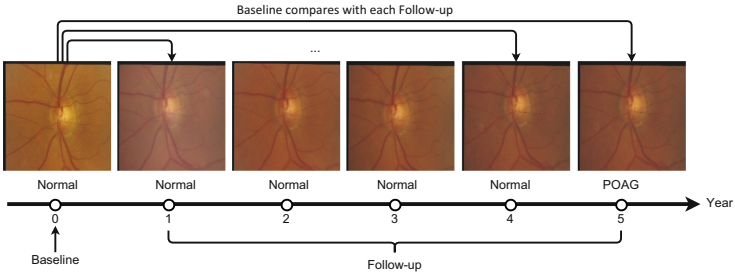


Fig. 1. Longitudinal fundus images of a patient.

Fundus photography provides a convenient and inexpensive way to record the optic nerve head structure, which is the gold standard for showing a classic glaucomatous appearance. Unfortunately, the low prevalence of glaucoma, the limited number of trained physicians, and the complex logistics of traditional screening programs are obstacles to timely screening for many patients [14]. Therefore, it is crucial to develop automatic models with high accuracy to assist clinicians in predicting future glaucoma event from fundus photographs, which can help many patients avoid blindness.

Recently, deep learning models have been successfully applied to biology and medicine [3,6–9,18,20,25,26]. In the ophthalmology domain, several models have been proposed to detect POAG from fundus photographs [2,4,16,17]. All of these studies predict the current glaucomatous status of a patient.

In this study, we seek to predict the probability of POAG onset from fundus photos. Such prediction may identify patients appropriate for early treatment. The Ocular Hypertension Treatment Study cohort (OHTS) [12], a large-scale clinical trial from 22 centers in the United States, includes longitudinal fundus photographs and disease assessment over a period of approximately 16 years. It provides an unprecedented opportunity to investigate POAG onset prediction using the dynamic fundus images.

To the best of our knowledge, only two previous works have focused on prediction of future POAG event [15,24]. However, the input to all these approaches is a single image, which might affect the performance of the model. In clinical practice, patients usually have follow-up visits to screen for glaucoma progression. Therefore, the glaucoma specialists compare the follow-up with the baseline image (the image taken at the first visit of a study) to trace the relevant feature, as shown in Fig. 1.

In this study, we used fundus images to predict an eye's progress to POAG (which may never occur) within specific inquired durations from the current visit.

The inquired duration was selected in advance (2-year or 5-year), and it was relative to the time when the image was taken, not to the time of the baseline visit. Unlike prior studies, for one eye, the inputs included one fundus image taken at the baseline (first visit), and one image was taken at the current visit (follow-up image). Therefore, our proposed method is suitable for screening patients during follow-up visits. We never need "future images" to screen people. The output was the probability that the time to POAG onset exceeds the inquired duration. To handle the pair of images, we proposed a novel Siamese network model with side output and additional convolution, called multi-scale multi-structure Siamese network (MMSNet), by comparing the differences between two input images. Different from previous Siamese work, MMSNet used the additional 2D features by convolution operation together with cosine similarity, instead of the cosine similarity only to measure the similarity between two outputs of the network instead of the cosine similarity between the two outputs. In addition, the MMSNet also consists of side output [19] to ease the vanishing gradient problems in training deep models and to drive the hidden layers for favoring discriminative features. To the best of our knowledge, it is the first time in the ophthalmology domain that two fundus images have been utilized and compared for automated glaucoma prediction before onset by Siamese networks.

Our work has the following contributions: (1) We proposed a model to compare the similarity between baseline and follow-up images. The model simulates the glaucoma screening process in clinical practice. To the best of our knowledge, no previous studies tackle the glaucoma prediction before onset problem this way. (2) We use the 2D features by convolution operation together with cosine similarity, instead of the cosine similarity only to measure the similarity between the outputs of two networks. (3) We incorporate the side-output to improve the performance further. The side-output can ease the vanishing gradient problems in training deep models and drive the hidden layers to favor discriminative features. (4) Our approach achieves superior POAG prediction 2-year and 5-year before onset results (93.12% and 95.07% in AUC) against several competitive baselines on the large, multi-institutional benchmarks.

2 Methods

2.1 MMSNet Architecture

MMSNet comprises two convolutional blocks that share the weight and two prediction blocks (Fig. 2). In the beginning, two fundus images x_1 and x_2 are passed through the convolutional neural network, DenseNet-201 [10], respectively. We used the output of last (F_{d4} and F_{d4n}) and second to last Dense Blocks (F_{d3} and F_{d3n}). For each output, we feed the outputs into the prediction block. Each prediction block has two paths to generate prediction results. For the first path, we got the feature embedding by average pooling, followed by cosine similarity to measure the similarity of feature embedding, which is a traditional Siamese network operation. We got the prediction result by sigmoid activation. Another path is the 2D block. We concatenated two outputs, followed by a 1×1 convolution, a batch normalization (BN), and rectified linear units (ReLU). In the end,

a global average pooling and a fully connected layer with sigmoid activation are attached. The average of the predicted values of the two paths is used as the final output of each prediction block, regarded as the side output.

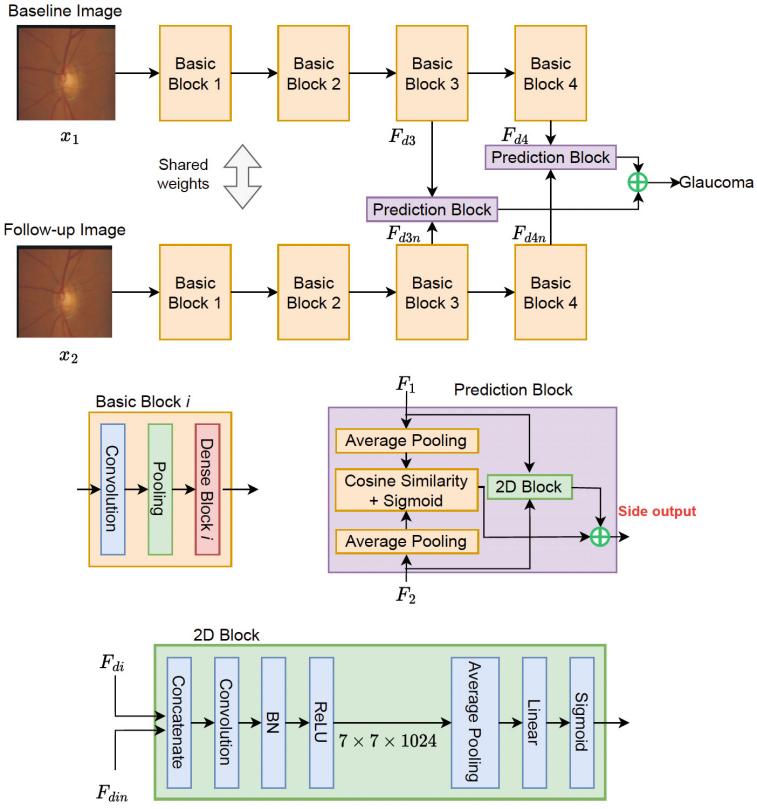


Fig. 2. The architecture of the proposed MMSNet.

2.2 Loss Function

In this study, we use binary cross-entropy as the loss function in the MMSNet. In addition, to overcome the severe class imbalance for the POAG classification, we apply the weighted cross-entropy, a commonly used loss function in classification. The adopted weighted cross-entropy was as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}[\beta y_n \log(\hat{y}_n(x_n, \theta_s)) + (1-\beta)(1-y_n)\log(1-\hat{y}_n(x_n, \theta_s))] \quad (1)$$

N is the number of training examples. β is the balancing factor between positive and negative samples. Here, we used inversely proportional to POAG frequency

in the training data. y_n is the observe value, $\hat{y}$ is the probability predicted by the classifier, and θ_s represents the parameters of the neural network.

The overall loss function is the average of the losses associated with the prediction from the two prediction blocks:

$$\mathcal{L}_s = \alpha\mathcal{L}^{(1)} + (1-\alpha)\mathcal{L}^{(2)}, \tag{2}$$

2.3 Data Augmentation

In this work, we sequentially apply the following augmentation techniques on the fly during training: (1) random rotation between 0° and 10°, (2) random translation: an image was translated randomly along the x- and y-axes by distances ranging from 0 to 10% of width or height of the image, and (3) random flipping.

3 Results

3.1 Datasets

In this study, we include one independent dataset (Table 1). This database is a large-scale, cross-sectional, longitudinal, and population-based study.

The dataset is obtained from the Ocular Hypertension Treatment Study (OHTS). OHTS is one of the largest longitudinal clinical trials in POAG (1,636 participants and 37,399 images) from 22 centers in the United States. This study does not need Institutional Review Board approval because it does not constitute human subjects research.

The participants in this dataset were selected according to both eligibility and exclusion criteria. The gold standard POAG labels were graded at the Optic Disc Reading Center. In brief, two masked certified readers were arranged to independently detect the optic disc deterioration. If there was a disagreement between two readers, a senior reader reviewed it in a masked fashion. The POAG diagnosis in a quality control sample of 86 eyes (50 normal eyes and 36 with progression) showed test-retest agreement at $\kappa = 0.70$ (95% confidence interval [CI], 0.55–0.85). More details of the reading center workflow have been described in [12]. For the OHTS dataset, we split the entire dataset randomly at the patient level. We take one group (20% of total subjects) as the hold-out test set and the remaining as the training set.

We compare the baseline image with each follow-up image and they compromise pairs separately. In each pair, the follow-up image and the baseline image come from the same eye. In this study, all eligible subjects are non-POAG at baseline. For 2-year POAG prediction before onset, in POAG pair, x_1 refers to the baseline image and x_2 refers to the follow-up image that coverted to POAG within two years. In normal pair, x_1 refers to the baseline image and x_2 refers to the follow-up image that did not convert to POAG within two years no matter whether it converts to POAG eventually. The definition of the POAG pair and the normal pair in the 5-year POAG prediction before onset can be derived similarly.

Table 1. Characteristics of the OHTS.

	Train	Val	Test
Participants	2,503	115	654
2-year prediction			
POAG	463	133	163
Normal	23,315	1,557	6,113
5-year prediction			
POAG	961	284	336
Normal	22,817	1,406	5,940

3.2 Evaluation Metrics

To evaluate the performance of POAG within a certain duration, we compute accuracy, sensitivity, specificity, and AUC (Area Under the ROC curve).

3.3 Experimental Settings

We first trained DenseNet-201 on POAG prediction using a single image as input. Then we initialized the subnets in the MMSNet using the DenseNet-201 and fine-tuned the entire network in an end-to-end manner.

All images are resized to $224 \times 224 \times 3$ as input of the proposed model. The models were implemented by Keras with a backend of Tensorflow. The proposed network was optimized using the Adam optimizer method [13]. The learning rate is 5×10^{-5} and α is 0.8. The experiments were performed on Intel Core i9-9960 X 16 cores processor and NVIDIA Quadro RTX 6000 GPU. The training time was 103 mins, and the testing time was 12 mins.

3.4 Results and Discussion

We compare our method with six models on POAG prediction on the OHTS dataset, including the DenseNet-201 with a single image as input, MobileV2 with a single image as input [24], the traditional Siamese network with an absolute difference, the traditional Siamese network with cosine similarity, MMSNet using the last DenseNet Block (MMSNet w/o side output), and MMSNet without 2D Block (MMSNet w/o 2D block).

2-Years POAG Prediction Before the Onset. Table 2 shows the performance comparison for 2-year POAG prediction before onset. Our model achieved the best results, with an accuracy of 0.9337, a sensitivity of 0.6485, a specificity of 0.9414, and an AUC of 0.9312. Compared to DenseNet-201, which is the best model among those baseline models with a single image as input, the proposed MMSNet possesses 27.90% higher accuracy, 28.70% higher specificity, and 22.04% higher AUC, let alone MobileV2 which also uses a single image as input

that is used in [24]. The result obtained by the Siamese network with cosine similarity is better than the results obtained by the DenseNet-201, indicating that the Siamese network, which imitates the clinical process (with multiple inputs), is more precise for 2-year POAG prediction before onset. As the results listed in row 3 and row 4, the Siamese network with cosine similarity works better than the Siamese network with an absolute difference.

Table 2. Comparisons on the test set of the OHTS dataset for 2-years POAG prediction before onset.

Method	Accuracy	Sensitivity	Specificity	AUC
DenseNet-201	0.6547	0.6667	0.6544	0.7108
MobileV2	0.8368	0.1697	0.8549	0.5114
Siamese network (absolute difference)	0.8634	0.4424	0.8748	0.8089
Siamese network (cosine similarity)	0.9423	0.4121	0.9566	0.8798
MMSNet w/o side output	0.9047	0.7515	0.9089	0.9085
MMSNet w/o 2D block	0.9132	0.6848	0.9193	0.8987
MMSNet	0.9337	0.6485	0.9414	0.9312

5-Year POAG Prediction Before the Onset. Table 3 compares the results of MMSNet with 6 state-of-the-art models on the OHTS dataset for POAG prediction 5-years before onset. Our model obtained the best results, with an accuracy of 0.9414, a sensitivity of 0.7530, a specificity of 0.9520, and an AUC of 0.9507. Compared to the baseline (DenseNet-201), MMSNet has higher accuracy (12.48% improvement), sensitivity (12.80% improvement), specificity (12.46% improvement), and AUC (14.90% improvement). Analogously, in this 5-year task, the result obtained by the Siamese network with cosine similarity is better than the results obtained by the DenseNet-201, indicating that the imitation of the clinical process (with multiple inputs) is superior to single-visit input. As the results listed in row 3 and row 4 of Table 3, the Siamese network with cosine similarity works better than the Siamese network with an absolute difference.

Table 3. Comparisons on the test set of the OHTS dataset for 5-years POAG prediction before onset.

Method	Accuracy	Sensitivity	Specificity	AUC
DenseNet-201	0.8166	0.625	0.8274	0.8017
MobileV2	0.5915	0.6667	0.5872	0.6799
Siamese network (absolute difference)	0.9312	0.5536	0.9525	0.9109
Siamese network (cosine similarity)	0.9433	0.6250	0.5412	0.9210
MMSNet w/o side output	0.9111	0.7292	0.9219	0.9238
MMSNet w/o 2D block	0.9348	0.6012	0.9535	0.9363
MMSNet	0.9414	0.7530	0.9520	0.9507

3.5 Ablation Studies

In this section, we conduct the ablation study to analyze the effect of the two components (1) using multi-scale features by side output; (2) using multi-structure features with the 2D block integrated into the proposed network. Note that MMSNet will reduce to the traditional Siamese network with cosine similarity by removing the side output and the 2D block. The results obtained by the Siamese network with cosine similarity are listed in the fourth row of Table 2 and Table 3. The 5 and 6 rows list the performance obtained by MMSNet without side output and MMSNet without 2D block, respectively. The results demonstrated that both the side output mechanism and 2D block utilize convolution to measure the similarity between two outputs could boost the performance of MMSNet. The last row lists the results obtained by the proposed method that consists of both side output and 2D block, and the results improve further.

4 Conclusions

In conclusion, this study proposed a new end-to-end deep learning network that simulates the process for automatic POAG prediction within a certain duration from fundus photographs. It is a first attempt to predict the POAG before onset by simulating the glaucoma screening process. The proposed network consists of a 2D block and side output. The 2D block via convolution operation utilizes the 2D features and cosine similarity to measure the similarity between two outputs. The side output drives the hidden layer for favoring discriminative features. One large dataset from multi-institutions was used to evaluate the proposed model. The results demonstrated that the proposed network performs well on POAG prediction before onset.

Acknowledgment. This work was also supported by awards from the National Eye Institute, the National Center on Minority Health and Health Disparities, National Institutes of Health (grants EY09341, EY09307), Horncrest Foundation, awards to the Department of Ophthalmology and Visual Sciences at Washington University, the NIH Vision Core Grant P30 EY 02687, Merck Research Laboratories, Pfizer, Inc., White House Station, New Jersey, and unrestricted grants from Research to Prevent Blindness, Inc., New York, NY.

References

1. Bourne, R.R., et al.: Causes of vision loss worldwide, 1990–2010: a systematic analysis. Lancet Glob. Health **1**(6), e339–e349 (2013)
2. Chen, X., Xu, Y., Wong, D.W.K., Wong, T.Y., Liu, J.: Glaucoma detection based on deep convolutional neural network. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 715–718. IEEE (2015)
3. Ching, T., et al.: Opportunities and obstacles for deep learning in biology and medicine. J. Roy. Soc. Interface **15**(141) (2018). https://doi.org/10.1098/rsif.2017.0387

4. Christopher, M., et al.: Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. Sci Rep. **8**(1), 1–13 (2018)
5. Doshi, V., Ying-Lai, M., Azen, S.P., Varma, R., Los Angeles Latino Eye Study Group, et al.: Sociodemographic, family history, and lifestyle risk factors for open-angle glaucoma and ocular hypertension: the Los Angeles Latino Eye Study. Ophthalmology **115**(4), 639–647 (2008)
6. Ehteshami Bejnordi, B., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**(22), 2199–2210 (2017). https://doi.org/10.1001/jama.2017.14585
7. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017). https://doi.org/10.1038/nature21056
8. Ghahramani, G.C., et al.: Multi-task deep learning-based survival analysis on the prognosis of late AMD using the longitudinal data in AREDS. medRxiv (2021)
9. Han, Y., et al.: Using radiomics as prior knowledge for thorax disease classification and localization in chest X-rays. In: AMIA Annual Symposium Proceedings, vol. 2021, p. 546. American Medical Informatics Association (2021)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
11. Jiang, X., Torres, M., Varma, R., Los Angeles Latino Eye Study Group, et al.: Variation in intraocular pressure and the risk of developing open-angle glaucoma: the Los Angeles Latino Eye Study. Am. J. Ophthalmol. **188**, 51–59 (2018)
12. Kass, M.A., et al.: The ocular hypertension treatment study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. Arch. Ophthalmol. **120**(6), 701–713 (2002)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Kolomeyer, N.N., et al.: Lessons learned from 2 large community-based glaucoma screening studies. J. Glaucoma **30**(10), 875–877 (2021)
15. Li, L., Wang, X., Xu, M., Liu, H., Chen, X.: DeepGF: glaucoma forecast using the sequential fundus images. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 626–635. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_60
16. Li, L., et al.: A large-scale database and a CNN model for attention-based glaucoma detection. IEEE Trans. Med. Imaging **39**(2), 413–424 (2019)
17. Li, Z., He, Y., Keel, S., Meng, W., Chang, R.T., He, M.: Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology **125**(8), 1199–1206 (2018)
18. Lin, M., Jiang, M., Zhao, M., Ukwatta, E., White, J.A., Chiu, B.: Cascaded triplanar autoencoder m-net for fully automatic segmentation of left ventricle myocardial scar from three-dimensional late gadolinium-enhanced MR images. IEEE J. Biomed. Health Inform. **26**(6), 2582–2593 (2022)
19. Lin, M., et al.: Fully automated segmentation of brain tumor from multiparametric MRI using 3D context deep supervised U-Net. Med. Phys. **48**, 4365–4374 (2021)
20. Lin, M., et al.: Artificial intelligence in tumor subregion analysis based on medical imaging: a review. J. Appl. Clin. Med. Phys. **22**(7), 10–26 (2021)
21. Quigley, H.A., Katz, J., Derick, R.J., Gilbert, D., Sommer, A.: An evaluation of optic disc and nerve fiber layer examinations in monitoring progression of early glaucoma damage. Ophthalmology **99**(1), 19–28 (1992)

22. Sommer, A., et al.: Racial differences in the cause-specific prevalence of blindness in east Baltimore. New Engl. J. Med. **325**(20), 1412–1417 (1991)
23. Tatham, A.J., Medeiros, F.A., Zangwill, L.M., Weinreb, R.N.: Strategies to improve early diagnosis in glaucoma. Prog. Brain Res. **221**, 103–133 (2015)
24. Thakur, A., Goldbaum, M., Yousefi, S.: Predicting glaucoma before onset using deep learning. Ophthalmol. Glaucoma **3**(4), 262–268 (2020)
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3462–3471. IEEE (2017). https://doi.org/10.1109/CVPR.2017.369
26. Wanyan, T., et al.: Supervised pretraining through contrastive categorical positive samplings to improve COVID-19 mortality prediction. In: Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 1–9 (2022)



HealNet - Self-supervised Acute Wound Heal-Stage Classification

Héctor Carrión[1(✉)], Mohammad Jafari[3], Hsin-Ya Yang[2], Roslyn Rivkah Isseroff[2], Marco Rolandi[1], Marcella Gomez[1], and Narges Norouzi[1]

[1] University of California, Santa Cruz, Santa Cruz, USA
hcarrion@ucsc.edu
[2] University of California, Davis, Davis, USA
[3] Columbus State University, Columbus, USA

Abstract. Identifying, tracking, and predicting wound heal-stage progression is a fundamental task towards proper diagnosis, effective treatment, facilitating healing, and reducing pain. Traditionally, a medical expert might observe a wound to determine the current healing state and recommend treatment. However, sourcing experts who can produce such a diagnosis solely from visual indicators can be difficult, time-consuming and expensive. In addition, lesions may take several weeks to undergo the healing process, demanding resources to monitor and diagnose continually. Automating this task can be challenging; datasets that follow wound progression from onset to maturation are small, rare, and often collected without computer vision in mind. To tackle these challenges, we introduce a self-supervised learning scheme composed of (a) learning embeddings of wound's temporal dynamics, (b) clustering for automatic stage discovery, and (c) fine-tuned classification. The proposed self-supervised and flexible learning framework is biologically inspired and trained on a small dataset with zero human labeling. The HealNet framework achieved high pre-text and downstream classification accuracy; when evaluated on held-out test data, HealNet achieved 97.7% pre-text accuracy and 90.62% heal-stage classification accuracy.

Keywords: Wound healing · Self-supervised medical image processing · Clustering

1 Introduction and Related Work

Wound healing is a variable, painful, and often prolonged process that can impact the quality of life for millions of people worldwide. Approximately 6.5 million patients present chronic wounds annually in the United States, requiring a treatment cost of over $25 billion per year [1]. In 2014, acute wounds resulted in 17.2

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_46.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 446–455, 2022.
https://doi.org/10.1007/978-3-031-21014-3_46

million hospital visits [2], while the odds of an acute wound requiring further care significantly increases with age, diabetes, and obesity [1]; thus a large portion of acute wounds will not resolve uneventfully and burden healthcare resources. An accessible and automated protocol to identify, track, and predict wound heal-stages could help tackle these challenges while reducing costs.

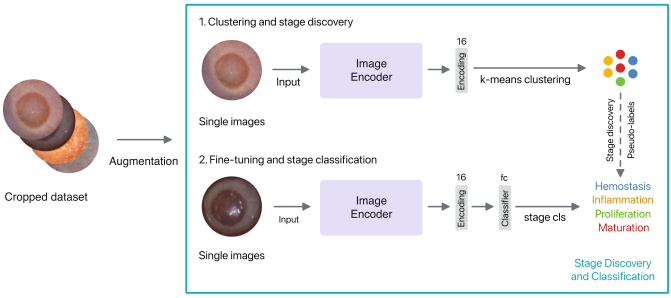


Fig. 1. (Step 1) The self-supervised network is used to generate 16-dimensional feature vectors. We cluster these feature vectors using k-means with $k = 4$. **(Step 2)** The new pseudo-labels (based on clusters) are used to assign each image a heal-stage class. The network is then re-tasked and fine-tuned using pseudo-labels, predicting the wound stage.

Biologically routine wound healing mechanics are four precise and programmed stages: hemostasis, inflammation, proliferation, and maturation [3]. Numerous chemical reactions and interactions between cells occur within these phases under and above the wound bed. However, the visual indicators that result from or relate to each stage are medically under-specified. Therefore, dermatologists are often limited to studying wound area changes or rely on mice wound re-epithelialization and histological analysis (which requires harvesting the wound-bed and thus sacrificing the mice) to understand wound-stage progression [4]. Invasiveness, the cost of long-running experiments, and the lack of qualified experts that can analyze image-based results contribute to the lack of large-scale, complete wound-stage progression image datasets.

Machine Learning (ML) for automated wound analysis has not been approached from the heal-stage angle. Methods exist for wound detection and size estimation [5], where obtaining ground-truth labels are relatively straightforward. Other classification systems aim to identify different characteristics of wounds or ulcers [6]. Acute wound linear regression models incorporate location, age, size, and infection features to generate time-to-heal estimates [7]. Recent lesion segmentation work has found significant inter-and-intra annotator variance even among experts [8], suggesting that precisely defining wound-stage visual indicators is not trivial.

Weakly-supervised and self-supervised learning approaches have shown the ability to improve the performance of tasks with little or no annotated data

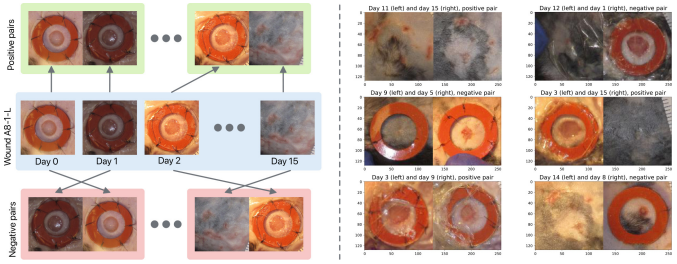


Fig. 2. (left side) Due to the nature of wound healing, images capturing progression naturally contain a temporal structure. We generate image pairs in the forwards (positive) and backward (negative) temporal directions; this serves as the input for our self-supervised temporal coherency training step. **(right side)** Randomly sampled image input pairs.

by optimizing a pre-text objective function that learns representations without (or with few) explicit human labels. For example, learning color from grayscale images (colorization) [9,10], shuffling and predicting image tile permutations (Jigsaw) [11], removing and re-generating image regions (inpainting) [12,13] rotation prediction (RotNet) [14] and classifying augmented positive and negative pairs (contrastive learning) [15,16]. In modalities like video and audio, the inherent temporal structure of the data can be exploited to create other proxy objective functions, for example shuffling video frames and predicting temporal order [17]. Further work has shown that these embeddings can be clustered and re-used as pseudo-labels to improve representation quality for weakly-supervised and self-supervised models [18].

Due to the nature of wound healing, capturing progression over time involves learning the inherent temporal structure. Thus, a proxy task is created to learn wound healing temporal dynamics. We study a small, unlabeled dataset containing daily mouse wound observations from onset to maturation. We leverage the dataset to generate wound image pairs in the temporal forward (positive) and backward (negative) directions. This step increases our dataset size by a factor of 15. The wound pairs are then used to train a network to determine the input pair's temporal validity and thus encoding features relevant to wound healing temporal dynamics. These embeddings are then clustered into four groups (one per heal-stage) to generate pseudo-labels without human intervention. Finally, a classifier is re-tasked and fine-tuned on pseudo-labels to predict the heal-stage of input images (Fig. 1).

2 Description of Data

The experiment for generating the image dataset used in this work is described in [19], from which we generate circular wound-only crops. The images contain wounds from eight mice (four young and four aged) imaged daily over sixteen days. Each individual had a wound inflicted on the left side and another on

the right side. The resulting dataset is 256 images (8 mice × 2 wounds × 16 days). Note that this dataset was not captured with computer vision in mind; consequently, we need to account for challenges such as blur, occlusion and illumination noise. To address this, data augmentation was used in training.

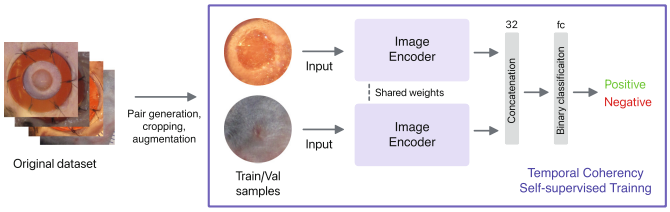


Fig. 3. Positive and negative wound pairs train the network. The image encoder forms a Siamese tuple configuration with shared parameters; each branch outputs a 16-dimensional feature vector. The feature vectors are concatenated and passed through a fully-connected binary classifier, outputting the input pair temporal validity.

Wound pairs have been created for the pre-text learning task in order to capture the temporal healing process. We generated pairs in the temporal forwards (positive) and backwards (negative) directions. Figure 2 illustrates the process of creating these pairs. We did not generate pairs containing same-day same-wound images. Thus, the resulting dataset contains 3,840 image pairs (16 wounds × 16 days × 15 pairs per day). In order to allow our validation and test sets to capture any differing dynamics between the two cohorts of young and aged mice, we randomly selected two young wounds and two aged wounds from the dataset; a couple of unique aged and young wounds is placed into the validation set and a second couple into the testing set. After this split, the resulting image pair dataset contains 2,880 training samples (75%), 480 validation, and testing samples (12.5% each). The single image dataset then contains 192 training samples (75%), 32 validation, and testing samples (12.5% each). We consistently keep the same validation and test wounds held out throughout the training pipeline and between our pre-text and downstream tasks. Finally, we cropped identical circles around the centroid of each wound image to ensure no signals other than the wound itself are contained in the training data.

3 Approach

The HealNet scheme involves a self-supervised learning step, followed by a clustering step, and a second learning step for wound-stage classification. Across all steps, the same image encoder neural network backbone is used and the trained weights are kept. Given the limited size of our training dataset, we chose DenseNet121 [20] as the HealNet backbone architecture. DenseNet is fully convolutional and thus includes fewer parameters than other commonly used feature

extractors while showing strong performance metrics. The first learning step revolves around self-supervised pre-text learning and thus includes a concatenation layer and a fully-connected binary classifier as the network head. The second learning step resembles a more traditional image classifier, replacing the previous head with a single fully-connected layer for heal-stage classification. Before training, DenseNet is initialized to ImageNet [21] weights.

3.1 Temporal Coherency Self-supervised Training

To predict the temporal validity of our image pairs and thus encode temporal dynamics, we define a tuple Siamese architecture [22] using two parallel stacks of identical feature extraction networks. Each stack shares weights and takes a single image from the pair as input while outputting a representation encoding (of size 16). We concatenate both encodings and feed them into a single fully-connected binary classification layer that determines whether the input pair is negative or positive. The loss function is binary cross-entropy and the optimizer is Adam [23] with default parameters. Figure 3 summarizes the temporal coherency training scheme.

3.2 Unsupervised Wound Stage Clustering

In order to identify and cluster wound healing stages, we use unsupervised k-means clustering. We apply unsupervised k-means clustering to the representations generated by our temporal coherency network to generate our clusters. We set $k = 4$ since medical literature suggests four main healing stages: hemostasis, inflammation, proliferation, and maturation. We use these cluster mappings to assign pseudo-labels to our dataset. Figure 1 illustrates the two steps of assigning pseudo-labels using k-mean clustering and subsequently using these pseudo-labels for heal-stage classification [18]. The question of how we specifically relate each cluster to each heal-stage is covered in Sect. 4.

3.3 Wound Stage Classification

As shown in Fig. 1, the final step of the pipeline is to re-task and fine-tune the pre-text network towards predicting the heal-stage of input images based on cluster assigned pseudo-labels. The re-tasking is done by adding a fully-connected layer with a softmax activation function and four output neurons (one per healing stage). The fine-tuning is accomplished by further training and updating the weights of the model using a low learning rate ($\eta = 0.001$) and the Adam [23] optimizer.

4 Results and Discussion

This section covers loss and accuracy scores for both learning steps, as well as evaluates the quality of cluster-based stage discovery from a cluster-member

appearance perspective, from a cluster-member wound-age statistics perspective and via agreement against human heal-stage annotators. Finally, we will cover the metrics achieved by the model on held-out test data at each learning step.

4.1 Temporal Coherency Self-supervised Evaluation

The self-supervised model is trained until convergence. Our results show the network is not overfitting, achieves high accuracy on the validation set, and can predict the temporal validity of the train, validation, and test sets. The prediction accuracy for the temporal validity task is 97.2%, 95.0%, and 97.7% on training, validation, and test sets, respectively, as compiled in Table 2.

4.2 Wound Healing Clustering Evaluation

We performed two-dimensional PCA in order to visualize the clusters along with their centroids. We can observe in Fig. 4 that our clusters are mostly close together, especially in the later stages. Given how wound progression is slowly evolving and the time-step between observations is equal to 1 day, this closeness and some overlap are expected. We can also observe how the hemostasis cluster contains the highest spread. This is likely because this cluster mostly contains day 0 images where the wound bed is uncovered and artificially illuminated and day 1 images where the wound bed is no longer artificially illuminated and covered with a protective plastic film, causing some distance between the embeddings. Due to the noisy nature of the dataset, some spread is also expected.

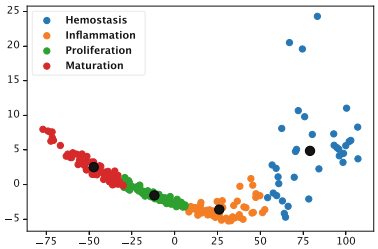


Fig. 4. K-means ($k = 4$) cluster visualization with centroids using 2-dimensional PCA.

To further evaluate the quality of the clusters, we looked into the images in each cluster (for young and aged mice separately). The median wound age (in days) and first and third quartile of wound age are calculated for each cluster. As can be seen from Fig. 5, images in each of the clusters contain visual cues for hemostasis (fresh and clearly defined wound-edge), inflammation (swelling of the wound edge, wet and shiny appearance), proliferation (dry and matte appearance with rough and uneven texture), and maturation (complete or near-complete wound closure). In addition, there is a clear temporal progression in the median wound day age that also differentiates our clusters. It would be interesting to

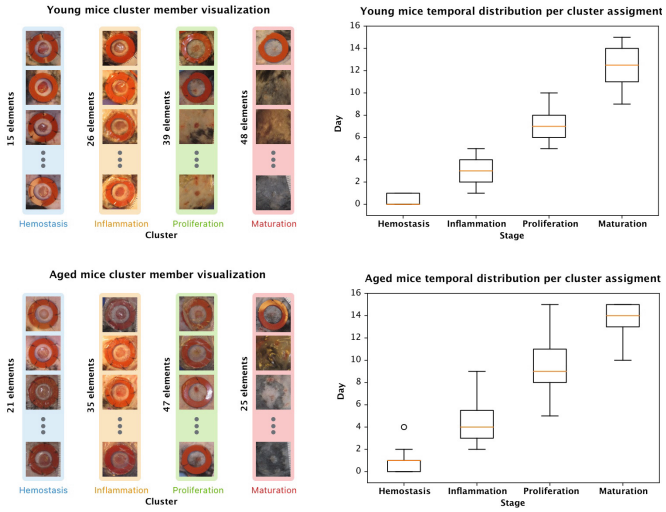


Fig. 5. (left) Cluster member examples split between young and aged mice and **(right)** Temporal distribution box plots. Note: here we show the original non-cropped images to verify the medical red rings (splints) have not biased the cluster members. We align the cluster groups with a heal-stage according to two metrics: visual appearances and wound age distribution metrics.

note that a differing wound healing rate between young and aged mice is encoded and shown in the median, first, and third quantiles of each cohort. Previous work on this dataset [19] also found a similar divergence in terms of wound-size progression between young and aged mice, aligning with these results. These two aspects are used in relating each cluster to a matching heal-stage. To validate these assignments, a blind survey of 5 human non-experts was performed where we found a top-1 agreement of over 80% between the self-supervised and human labels. Table 1 outlines detailed statistics on cluster temporal distribution and Table 2 describes human annotator agreement per datasplit.

Table 1. Cluster group wound-age (experimental days 0–15) statistics.

Age group	Cluster	Q1	Median (Q2)	Q3	Average
Young	Cluster 1	0.0	0.0	1.0	0.5
	Cluster 2	2.0	3.0	4.0	3.0
	Cluster 3	6.0	7.0	8.0	7.1
	Cluster 4	11.0	12.5	14.0	12.5
Aged	Cluster 1	0.0	1.0	1.0	1.0
	Cluster 2	3.0	4.0	5.5	4.5
	Cluster 3	8.0	9.0	11.0	9.6
	Cluster 4	13.0	14.0	15.0	13.4

4.3 Wound Stage Classification

The wound stage classification model is trained until convergence. While this task contains fewer training samples (15× less), the loss curves show the network is not overfitting, achieves relatively high accuracy on the validation set, and can predict the heal-stage of training, validation, and test sets (Table 3). The accuracy for the wound stage classification task is 95.8%, 87.5%, and 90.6% on training, validation, and test sets, respectively. Finally, we trained an identical architecture as a straight-forward CNN Classifier (without the pre-text task) utilizing the collected human survey labels in order to compare performance. Under these conditions the network achieved a lower accuracy of 83.3%, 81.3%, and 78.1% on training, validation, and test sets, respectively. Table 3 summarizes the loss and accuracy scores described above.

Table 2. Human annotator vs. automatic pseudo-label agreement.

Data split	Agreement (%)
Train	80.7
Validation	71.9
Test	87.5
Overall	80.5

Table 3. Performance metrics across both temporal coherency and wound stage classification tasks for training, validation, and test sets.

Task	Train acc.	Val. acc.	Test acc.
Supervised classifier (baseline)	83.3%	81.3%	78.1%
Temporal coherency (pre-text)	97.2%	95.0%	97.7%
Wound Stage classification (downstream)	**95.8%**	**87.5%**	**90.6%**

5 Conclusion

HealNet is a self-supervised learning scheme for automatic wound stage discovery and classification. It forms a novel combination of pre-text temporal feature extraction, k-means clustering, and pseudo-label re-tasking and fine-tuning. The results demonstrate the effectiveness and performance of the presented method with zero human-assisted data annotation. Considering the cost-effectiveness of the approach compared to medical experts' continuous diagnosis, the developed pipeline is a promising tool for augmenting, accelerating, and improving traditional human diagnosis. Given that there exists a high chance for expert disagreement in the medical field, a self-supervised approach could aid in decision-making or in compiling a true 'golden-master' list of visual indicators relating to

the healing process, the healing stages and the temporal patterns between them. The proposed HealNet framework can also be used by the scientific community to discover obscure and unseen patterns in their unlabeled data at low cost.

Acknowledgements. This research was sponsored by the Defense Advanced Research Projects Agency (DARPA), Army Research Office, under Cooperative Agreement No. W911NF-18-2-0104, and the Department of Interior, Award No. D20AC00003.

References

1. Sen, C.K., et al.: Human skin wounds: a major and snowballing threat to public health and the economy. Wound Repair Regener. **17**(6), 763–771 (2009)
2. Steiner, C.A., Karaca, Z., Moore, B.J., Imshaug, M.C., Pickens, G.: Surgeries in hospital-based ambulatory surgery and hospital inpatient settings, 2014: statistical brief# 223 (2020)
3. Guo, S., DiPietro, L.A.: Factors affecting wound healing. J. Dent. Res. **89**(3), 219–229 (2010)
4. Bagood, M.D., et al.: Re-examining the paradigm of impaired healing in the aged murine excision wound model. J. Investig. Dermatol. **141**(4S), 1071–1075 (2021)
5. Chino, D.Y., Scabora, L.C., Cazzolato, M.T., Jorge, A.E., Traina-Jr, C., Traina, A.J.: Segmenting skin ulcers and measuring the wound area using deep convolutional networks. Comput. Methods Program. Biomed. **191**, 105376 (2020)
6. Oyibo, S.O., Jude, E.B., Tarawneh, I., Nguyen, H.C., Harkless, L.B., Boulton, A.J.: A comparison of two diabetic foot ulcer classification systems: the Wagner and the university of Texas wound classification systems. Diabetes Care **24**(1), 84–88 (2001)
7. Ubbink, D.T., Lindeboom, R., Eskes, A.M., Brull, H., Legemate, D.A., Vermeulen, H.: Predicting complex acute wound healing in patients from a wound expertise centre registry: a prognostic study. Int. Wound J. **12**(5), 531–536 (2015)
8. Mirikharaji, Z., Abhishek, K., Izadi, S., Hamarneh, G.: D-LEMA: deep learning ensembles from multiple annotations-application to skin lesion segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1837–1846 (2021)
9. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 577–593. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_35
10. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 649–666. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_40
11. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
12. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)

13. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
14. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
15. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
16. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
17. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 527–544. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_32
18. Yan, X., Misra, I., Gupta, A., Ghadiyaram, D., Mahajan, D.: ClusterFit: improving generalization of visual representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6509–6518 (2020)
19. Carrión, H., Jafari, M., Bagood, M.D., Yang, H.-Y., Isseroff, R.R., Gomez, M.: Automatic wound detection and size estimation using deep learning algorithm. PLOS Comput. Biol. **18**, e1009852 (2022)
20. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
21. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
22. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "Siamese" time delay neural network. In: Advances in Neural Information Processing Systems, vol. 6 (1993)
23. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)



Federated Tumor Segmentation with Patch-Wise Deep Learning Model

Yuqiao Yang[1,2,3(✉)], Ze Jin[1,2,3], and Kenji Suzuki[1,2,3]

[1] BioMedical Artificial Intelligence (BMAI) Research Unit, Tokyo Institute of Technology, Tokyo, Japan
yang.y.aq@m.titech.ac.jp
[2] Laboratory for Future Interdisciplinary Research of Science and Technology (FIRST), Tokyo Institute of Technology, Tokyo, Japan
[3] Department of Information and Communications Engineering, Tokyo Institute of Technology, Tokyo, Japan

Abstract. A chief challenge of deep learning in computer-aided diagnosis is to collect a large heterogeneous dataset from multiple hospitals for constructing a robust deep learning model, because of hospitals' strict rules for sharing sensitive medical data. Federated learning (FL) allows collaborative learning among hospitals without sharing data to prevent a leakage of private information. FL, however, has issues of high computational demands and data shortage in local hospitals. To address these issues, we developed an FL framework coupled with a patch-wise deep learning model, a massive-training artificial neural network (MTANN), in tumor segmentation in CT. We performed experiments on the proposed MTANN-based federated tumor segmentation with a small-sized multisite liver tumor dataset. Our model achieved a Dice of 0.712 comparable to 0.723 by the gold-standard centralized training model, which was higher than 0.592 ($P < 0.05$) by the state-of-the-art Res-U-Net model. Furthermore, the proposed framework required a training time of 0.25 hours that was less than that of 5 hours by the Res-U-Net on a GPU server.

Keywords: Federated learning · Image segmentation · Massive-training artificial neural networks

1 Introduction

Deep learning has become a promising approach to computer-aided diagnosis with medical images for its diverse ability to perform diagnostic tasks at high performance levels [1, 2]. However, deep learning requires a large quantity of training data collected from multiple hospitals for learning the heterogeneity among patients in different hospitals. In other words, a trained deep learning model would have low generalizability when the model is trained with a dataset only from single hospital. Thus, a chief challenge of deep learning in computer-aided diagnosis is to collect large datasets of diverse patients from heterogeneous sources, because the hospitals have strict rules for sharing sensitive medical data.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 456–465, 2022.
https://doi.org/10.1007/978-3-031-21014-3_47

Recently, federated learning (FL) has been proposed [3]; and it allows collaborative learning among hospitals without sharing data to prevent a leakage of private information. In FL, each deep learning model is trained on-site at each local hospital with its local medical data, and only intermediate model training updates are transferred to a central server without sharing the local data. Thus, sensitive medical data with patient identifiable information is stored in each hospital and not exchanged or transferred. Researchers have used the FL structure in medical diagnosis tasks such as brain tumor segmentation [4], diabetic retinopathy [5], and electronic health records [6].

FL has great potentials in computer-aided diagnosis. However, there are three major issues with FL. The first one is small-sized training data in each local hospital. The performance of a deep learning model at a local hospital becomes low because of a shortage of local medical data. The second one is a demand for high computational power for on-site training of a local deep learning model at each hospital. Each hospital needs a high-end graphics processing unit (GPU) sever. The third one is heavy network loads during the collaborative learning among hospitals, as client models at local hospitals have to be synchronized frequently by transmitting model weights.

To address these issues, we developed a novel FL framework coupled with a patch-wise deep learning model, a massive-training artificial neural network (MTANN) [7, 8]in tumor segmentation in hepatic CT [8]. Compared to convolutional neural network (CNN)-based deep learning models, the MTANN model has (1) a small-data training characteristic, i.e., training with a small-sized data set, (2) training at a low computational cost, and (3) small-sized model weights [9]. In our FL framework, we used a federated averaging algorithm [10] to aggregate the MTANN model updates from each local hospital without sharing data. The training framework is shown in Fig. 1.

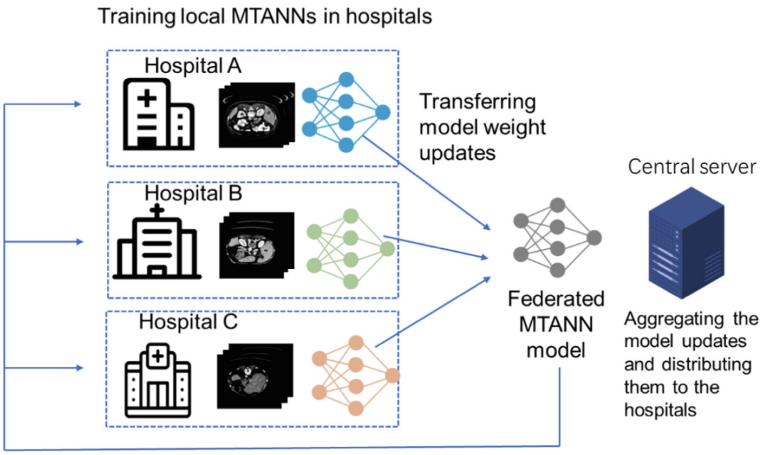


Fig. 1. The structure of our FL framework coupled with MTANNs

With our proposed framework, the training data size and computational power at local hospitals and network loads can be reduced, providing better collaborative learning among hospitals without data sharing. To evaluate our framework, we performed

experiments on the proposed MTANN-based federated tumor segmentation with a small-sized multisite liver tumor dataset. We will reveal the advantages of the MTANN FL framework through experiments.

2 Related Work

Federated Learning [2]**.** FL is a collaborative learning protocol in which several parties jointly train a deep learning model. Training occurs in rounds during which a central server sends the model to the local clients. Local training occurs for a number of iterations, after which the updated models are aggregated by federated averaging [5]. FL has a great potential in medical image diagnosis as it provides a way to train a model on multi-institutional collaboration without sharing the actual patient data. Sheller et al. [11] used FL for multi-institutional collaboration in brain tumor segmentation. They applied U-net [12] in an FL setting using only federated averaging algorithms. Li W et al. [13] offered differential-privacy techniques to protect the data during the weight transfer process in FL. According to the experimental results, there was a trade-off between performance and the privacy of the transferred weights.

Tumor Segmentation: In recent years, deep learning models were employed for tumor segmentation tasks. The CNN-based models have achieved great success in many semantic segmentation and object recognition problems in the computer vision community. Researchers began to utilize various CNNs for learning feature representations in the application of liver and lesion segmentations. Ben-Cohen et al. [14] proposed a cascaded FCN architecture and dense 3D conditional random fields to segment liver and liver lesions automatically. Meanwhile, Sun et al. [15] designed a multi-channel FCN to segment liver tumors from CT images, where the probability maps were generated by the feature fusion from different channels.

Sato et al. [8] introduced a 3D MTANN combined with a Hessian ellipse enhancer for liver tumor segmentation. The 3D MTANN has a substantial advantage of achieving high performance by training with a small number of cases. They have achieved a high-performance model with a small dataset. It has been proved that the MTANN architecture allows the model to achieve a high performance through training with a small amount of medical image data for suppressing ribs in chest radiographs [16], reduction of quantum noise in low-dose x-ray images[17] and semantic organ segmentation in x-ray images [18].

3 Method

3.1 Massive-Training Artificial Neural Network

The MTANN is a patch-wise supervised deep learning model that directly operates on voxel data. The MTANN was proposed by Suzuki et al. by extending supervised nonlinear image-processing techniques, called “neural filters” [19] and “neural edge enhancers” [20]. The input to an MTANN is voxel values in a sub-volume (i.e., a 3D image patch), V_s, extracted from 3D input images. The pipeline of training of an MTANN is presented in Fig. 2.

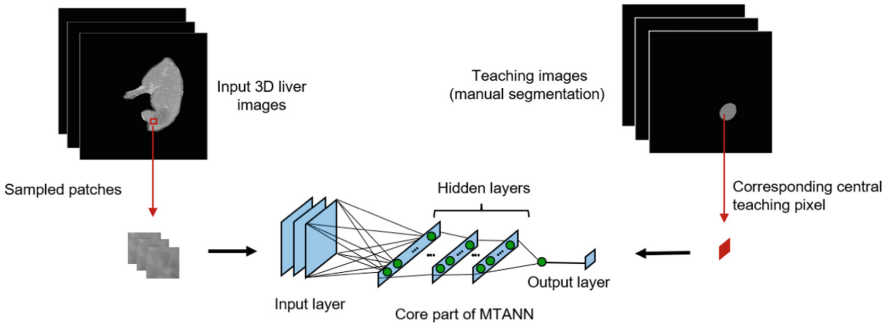


Fig. 2. Training of an MTANN in liver tumor segmentation in CT images

3D image patches are sampled and extracted from tumors and the areas around the tumors in 3D liver images. Each extracted patch has a corresponding label which is the central pixel value of the patch in the teaching image (i.e., "gold-standard" manual segmentation of a liver tumor by a radiologist). Then a neural network model with multiple fully connected layers is applied for patch-based classification to accomplish the segmentation task. The output $O(x, y, z)$ of the MTANN is the probability of being a tumor, and is presented by

$$O(x, y, z) = NN\{I(x - i, y - j, z - k)\} \tag{1}$$

where $i, j, k \in V_s$, and $NN(\cdot)$ is the neural network model of the MTANN, and $I(x, y, z)$ represents a voxel value of the 3D input volume.

The binary cross-entropy is applied as the loss function between the teaching image and the predicted likelihood map of being a tumor. Once the MTANN is trained, the whole 3D image is computed through shifting of the patch-wise window to obtain a likelihood map of being a tumor, in a convolutional manner across the entire input image.

3.2 Federated Learning Coupled with MTANN

Federated averaging [10] algorithm was implemented with the above 3D MTANN model to build our proposed FL framework on a client-server architecture. In this configuration, a centralized server maintains a global MTANN model and coordinates local stochastic gradient descent (SGD) updates of clients. The whole procedure is presented in Fig. 1. Firstly, a 3D MTANN model was initialized and broadcasted to all the clients (participated medical institutions/hospitals) at the start of training. Each client runs a client-side training algorithm on their local medical data and updates client MTANN models. When client-side training is finished, the model updates are sent to the server from each client. The server aggregates model updates from the clients with a server-side model aggregation algorithm. Then the server distributes the new model weights to clients to start the next round of federated training.

3.2.1 Client-Side Model Training

In each federated round t, patches p_k are sampled from the local client k's dataset. The local model is initialized by reading global model parameters $W^{(t)}$ from the server and is updated to $W_k^{(l,t)}$ by running multiple iterations of SGD (stochastic gradient descent), and l stands for a local iteration number. After running a fixed number of iterations $N^{(local)}$, a model difference $\Delta W^{(N^{(local)},t)}$ is transmitted to the aggregation server.

3.2.2 Server-Side Model Aggregation

The server receives updates from all clients at each federated round. Different clients may have various numbers of patients and training iterations. As the MTANN uses sampled patches instead of all images to train, it is essential to require the number of sampled patches and the number of local training iterations to aggregate the local models. The aggregation method is represented by

$$W^t \leftarrow W^{t-1} + \sum_{k=1}^{K} \frac{1}{\sum_{k=1}^{K} N_{pk} N_k^{(local)}} N_{pk} N_k^{(local)} \Delta W_k^{(N^{(local)},t)} \tag{2}$$

where N_{pk} stands for the number of sampled patches in client k, $N_k^{(local)}$ stands for the number of training iterations in client (hospital) k, and $\Delta W_k^{(N^{(local)},t)}$ stands for the model difference in client k.

4 Experiments Setting

This section describes the experimental setup, including dataset preparation, the common hyper-parameters used for the FL system, and methods for comparisons.

4.1 Materials

The database used in the experiments contains dynamic contrast-enhanced CT scans of 42 patients with 194 liver tumors acquired from the Liver Segmentation Benchmark (LiTS) database [21]. Each slice of the CT volumes has 512 × 512 pixels, with in-plane pixel sizes of 0.60–1.0 mm and slice thickness of 0.70–2.0 mm. The liver tumor region in CT was segmented by trained radiologists and then verified by three experienced radiologists.

4.2 Data Preparation

The datasets with 42 patients were divided into 18 training and 24 testing cases. Training cases were used to build our multi-site training datasets to train the model in the FL framework. To simulate different multi-site datasets, the images were randomly divided into three groups and six groups. The 24 testing cases were used for evaluating the performance of the model.

4.3 Federated Model Setup

Federated learning consists of two critical hyper-parameters. The number of federated rounds and local iterations. Federated round stands for a whole iteration of the FL process. Local iterations stand for the training iterations in local clients during the client-side model training. A local epoch is defined as that every client updates its local training examples exactly once. The MTANN used in the experiments has three fully connected layers with sigmoid activation functions. The shape of the input patches is a $5 \times 5 \times 5$ cube. Hessian enhanced images [8] were introduced as input images together with the original CT images. The number of units in the whole structure is 250, 50, 25, and 1. The number of federated rounds was set to 100, with 500 local iterations per federated round. The proposed model was trained with the binary cross-entropy, and a learning rate of 0.0005. For comparison of model convergences, the data-centralized baseline model was trained with 50,000 iterations.

4.4 Comparative Segmentation Method

One of the state-of-the-art deep-learning methods (Res-U-Net [22]) was used for comparison in the experiment. The Res-U-Net was built with residual units and batch normalization layers on the structure of U-Net to construct a deeper model. A Dice coefficient was used to evaluate the segmentation result. A Dice coefficient is a statistic used to measure the similarity of two regions.

5 Result and Discussion

5.1 Result of Tumor Segmentation

Table 1. Comparisons of the MTANN model with the state-of-the-art Res-U-Net model in FL training

# of hospitals	# of patients per hospital	Model	Dice
Centralized	18	MTANN with centralized data	0.723
		Res-U-Net with centralized data	0.653
3	6	**Federated MTANN**	**0.712**
		Federated Res-U-Net	0.592
6	3	**Federated MTANN**	**0.714**
		Federated Res-U-Net	0.532

The MTANN framework obtained average Dice scores of 0.712 (in the 3 hospitals scenario) and 0.714 (in the 6 hospitals scenario) which were comparable ($P > 0.90$) to 0.723 by the gold-standard centralized training model. The MTANN Dice results were higher than a Dice score of 0.592 ($P < 0.05$) by the state-of-the-art Res-U-Net model,

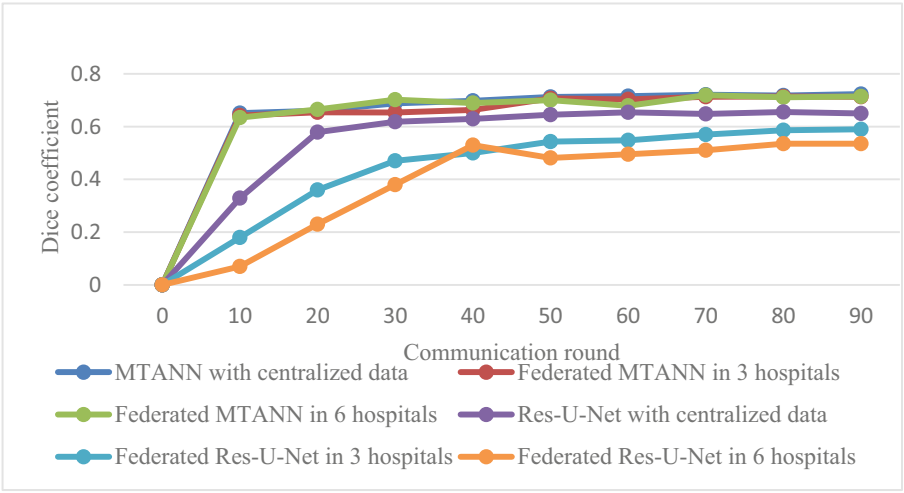


Fig. 3. Training curves of the MTANN and Res-U-Net in different distributed datasets

as shown in Table 1. The results showed that the proposed FL framework was able to achieve an equivalent segmentation performance without sharing clients' medical data. The performance (a Dice score of 0.714) of the MTANN was relatively robust even though the data were divided into six groups (meaning a small dataset in each local client). Compared to the robust performance of the MTANN, the 3D Res-U-Net had a worse performance as the number of the dataset groups (hospitals) increased. Thus, the 3D Res-U-Net model was more vulnerable with a small dataset. In addition, the MTANN models were faster to be trained, as shown in the training curves in Fig. 3. Our proposed framework had a better performance in all learning situations, compared to the Res-U-Net.

5.2 Computational Demands

With the MTANN, we only trained a relatively small model with patches (as opposed to the entire images of patients) in training. This training assignment allows clients to train a model with a low computational power (e.g., on an ordinary CPU PC). That would make FL more accessible to small practitioners/hospitals. Comparisons of computation demands of the MTANN and Res-U-Net are shown in Table 2.

The MTANN has much fewer parameters than does the Res-U-Net. The proposed MTANN framework required a training time of 0.25 or 3 hours that was less than 5 or 100 hours by the Res-U-Net on a GPU server or a CPU, respectively.

The training in the experiments was performed on a GPU server (CPU: Xeon E5–2698 v4, 2.20 GHz with 256 G RAM; GPU: Tesla V100-DGXS, Nvidia). In addition, the number of weights of the MTANN model is much smaller (72.2 KB in the file size) than that of the Res-U-Net (36.3 MB). Small size model weights reduce network traffic/loads during the model weights transmissions among hospitals. The MTANN has the following advantages over existing methods in federated tumor segmentation.

Table 2. Comparisons of computational demands

Model	# of Parameters	Training time on GPU	Training time on CPU	Size of model weights
MTANN	17,550	0.25 hours	3 hours	72.2 KB
Res-U-Net	9,498,260	5 hours	100 hours	36.3 MB

- **Better data protection:** A major privacy concern of FL is the gradient attack [23] to the server or clients, which tries to recover the original images from leaked model weights. The MTANN only uses randomly-sampled patches for the model training, which makes it difficult for the attacker to recover the original images.
- **Robust with a small dataset**: The MTANN consists of a unique pipeline that only trains a relatively small neural network model with the patches sampled from the input images. Focusing on the sampled patches can help reduce the variations of input image data. The model is robust because of the less diversity of the sampled patches compared to entire patients' images from different hospitals.
- **Low computational demands:** Since the MTANN is a relatively small model trained with a small training data of sampled patches (as opposed to the entire images of patients), a computational cost for training is very low, which allows local hospitals to join the FL with an ordinary CPU machine (as opposed to an expensive high-end GPU server). These low local computational demands make our FL framework more accessible in practical applications in hospitals.
- **Low network loads:** Since the size of MTANN model weights is a relatively small, the network traffic/loads are less heavy. The low network loads are desirable in FL.

6 Conclusion

We developed a new framework of FL coupled with patch-wise MTANN models to provide an accessible, robust FL model trained collaboratively among hospitals without sharing sensitive medical data. The performance of our FL framework with MTANN was equivalent to the reference-standard centralized training model and significantly higher than that of the state-of-the-art Res-U-Net model. Furthermore, our MTANN-based FL framework required a low computational demand; and thus, it provided a cost-efficient FL solution at local hospitals.

Acknowledgements. This research is based on results obtained from a project commissioned by the NEDO. The authors are grateful to all members of the BMAI at Tokyo Institute of Technology for their valuable discussions.

References

1. Litjens, G., Kooi, T., Bejnordi, B.E., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)

2. Suzuki, K.: Overview of deep learning in medical imaging. Radiol. Phys. Technol. **10**(3), 257–273 (2017). https://doi.org/10.1007/s12194-017-0406-5
3. Konečný, J., McMahan, B., Ramage, D.: Federated optimization distributed optimization beyond the datacenter (2015)
4. Ziller, A., Usynin, D., Remerscheid, N., et al.: Differentially private federated deep learning for multi-site medical image segmentation (2021)
5. Chang, K., Balachandar, N., Lam, C., et al.: Distributed deep learning networks among institutions for medical imaging. J. Am. Med. Inform. Assoc. **25**, 945–954 (2018). https://doi.org/10.1093/jamia/ocy017
6. Pfohl, S.R., Dai, A.M., Heller, K.: Federated and differentially private learning for electronic health records (2019)
7. Suzuki, K., Li, F., Sone, S., Doi, K.: Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network. IEEE Trans. Med. Imaging **24**, 1138–1150 (2005). https://doi.org/10.1109/TMI.2005.852048
8. Sato, M., Jin, Z., Suzuki, K.: Semantic segmentation of liver tumor in contrast-enhanced hepatic CT by using deep learning with hessian-based enhancer with small training dataset size. In: Proceedings - International Symposium on Biomedical Imaging. IEEE Computer Society, pp 34–37 (2021)
9. Tajbakhsh, N., Suzuki, K.: Comparing two classes of end-to-end machine-learning models in lung nodule detection and classification: MTANNs vs CNNs. Pattern Recognit. **63**, 476–486 (2017). https://doi.org/10.1016/j.patcog.2016.09.029
10. Brendan McMahan, H., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data (2017)
11. Sheller, M.J., Reina, G.A., Edwards, B., et al.: Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 92–104. Springer Verlag (2019)
12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 234–241. Springer Verlag (2015)
13. Li, W., Milletarì, F., Xu, D., et al.: Privacy-preserving federated brain tumour segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp 133–141. Springer (2019)
14. Ben-Cohen, A., Diamant, I.: Fully convolutional network for liver segmentation and lesions detection. Springer International Publishing, Cham (2016)
15. Sun, C., Guo, S., Zhang, H., et al.: Automatic segmentation of liver tumors from multiphase contrast-enhanced CT images based on FCNs. Artif. Intell. Med. **83**, 58–66 (2017). https://doi.org/10.1016/j.artmed.2017.03.008
16. Suzuki, K., Abe, H., MacMahon, H., Doi, K.: Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network (MTANN). IEEE Trans. Med. Imaging **25**, 406–416 (2006). https://doi.org/10.1109/TMI.2006.871549
17. Suzuki, K., Horiba, I., Sugie, N., Nanki, M.: Neural filter with selection of input features and its application to image quality improvement of medical image sequences. IEICE Trans. Inf. Syst. **85**, 1710–1718 (2002)
18. Suzuki, K., Horiba, I., Sugie, N., Nanki, M.: Extraction of left ventricular contours from left ventriculograms by means of a neural edge detector. IEEE Trans. Med. Imaging **23**, 330–339 (2004). https://doi.org/10.1109/TMI.2004.824238

19. Suzuki, K., Horiba, I., Sugie, N.: Efficient approximation of neural filters for removing quantum noise from images. IEEE Trans. Signal Process. **50**, 1787–1799 (2002). https://doi.org/10.1109/TSP.2002.1011218
20. Suzuki, K., Horiba, I., Sugie, N.: Neural edge enhancer for supervised edge enhancement from noisy images. IEEE Trans. Pattern Anal. Mach. Intell. **25**, 1582–1596 (2003). https://doi.org/10.1109/TPAMI.2003.1251151
21. Bilic, P., Christ, P.F., Vorontsov, E., et al.: The liver tumor segmentation benchmark (LiTS) (2019)
22. Dong nan da xue, Institute of Electrical and Electronics Engineers. 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP 2019): July 19–21, 2019, Southeast University (Wuxi), China (2019)
23. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the ACM Conference on Computer and Communications Security. Association for Computing Machinery, pp 1322–1333 (2015)



Multi-scale and Focal Region Based Deep Learning Network for Fine Brain Parcellation

Yuyan Ge[1,2], Zhenyu Tang[3], Lei Ma[2], Caiwen Jiang[2], Feng Shi[4], Shaoyi Du[1,5(✉)], and Dinggang Shen[2,4(✉)]

[1] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Shaanxi 710049, China
dushaoyi@xjtu.edu.cn
[2] School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China
dgshen@shanghaitech.edu.cn
[3] School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[4] Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200030, China
[5] Shunan Academy of Artificial Intelligence, Ningbo 315000, Zhejiang, China

Abstract. Brain parcellation plays an important role in neurodegenerative disease diagnosis and brain network analysis. One of the big challenges in brain parcellation is lack of clear anatomical boundary between different brain regions. As a result, for the task involving a large number of brain regions, i.e., during fine brain parcellation, the parcellation accuracy could be significantly degraded. Unfortunately, few studies focused on this issue. To this end, we propose a novel multi-scale deep brain parcellation network. Specifically, different scales of brain regions, i.e., global, middle and fine scales, are defined. From global to fine scales, brain regions are gradually subdivided and refined. The proposed deep network performs brain parcellation at each scale simultaneously (multi-task), where parcellation at fine scale is under the constraint of large scales. In addition, we also present a new focal region based auxiliary network, which focuses on the brain regions difficult to be parcellated at fine scale. The final parcellation results are obtained by merging the outputs of the brain parcellation backbone at all scales and the focal region based auxiliary network. Comparison and ablation experiments are conducted on a multi-center clinical brain MRI dataset of 267 subjects with 101 brain regions. Experimental results demonstrate that the proposed approach outperforms the state-of-the-art methods under comparison.

Keywords: Brain parcellation · Multi-task · Multi-scale labels · Segmentation · Deep learning

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-21014-3_48.

© Springer Nature Switzerland AG 2022
C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 466–475, 2022.
https://doi.org/10.1007/978-3-031-21014-3_48

1 Introduction

Brain parcellation plays an important role in neurodegenerative disease diagnosis, brain network construction, and brain quantitative analysis. It divides the brain into non-overlapping brain regions using structural magnetic resonance (MR) brain images. Since the complicated brain structure, manual brain parcellation is a professional and time-consuming task, and automatic brain parcellation is highly desired.

Traditional multi-atlas based brain parcellation (MAP) [4,12,14] was the most popular method for automatic brain parcellation in the past. It registers brain atlases to the target brain image, by which manually labeled brain regions in the atlases can be propagated to the target brain image. However, MAP depends on image registration, which is computationally inefficient and sensitive to image noise. Currently, many deep-learning based methods have been proposed to replace core components (i.e., image registration, and label fusion) of MAP. Currently, some studies focused on deep-learning based brain parcellation (DLP) methods [1,2,5,6,10,11,13] to achieve end-to-end brain parcellation. As benefiting from deep-learning, DLP methods have shown high computational efficiency and robustness compared with MAP.

Although existing DLP methods have shown promising results, few studies work with fine brain parcellation, where a large number of brain regions are required to be parcellated. It is known that the definition of brain regions is partially based on anatomical structure, brain function, metabolism, neurotransmitters, and gene expression, brain parcellation using pure image is difficult as there are often no clear anatomical boundaries between different brain regions. As a result, for brain parcellation involving a large number of brain regions, i.e., during fine brain parcellation, significantly degraded performance is inevitable.

To solve the aforementioned issue of fine brain parcellation, in [2,5,6], multi-model based methods have been proposed, where each model is built at different location or view of the brain. However, with the increased number of networks, e.g., 8 or 27 [6], or even 250 3D networks [2], computational efficiency could be decreased in both training and inference phases. In [15], a coarse-to-fine method was presented, where a network using low-resolution images was first trained to get the coarse location of one brain region, and then high-resolution images of cropped patches were used to train a fine network to get the final brain region. However, like the multi-model based methods, the coarse-to-fine methods suffer from low efficiency, because for each brain region, a fine network has to be built. In addition, both kinds of methods treat each brain region independently, ignoring their correlations.

In this paper, we propose a novel multi-scale and focal region based framework to accomplish accurate and efficient fine brain parcellation. Specifically, a multi-scale brain parcellation network is proposed to perform the parcellation in multi-scale brain regions (i.e., global, middle and fine scales). From global scale to fine scale, brain regions are sub-divided and refined in a hierarchical manner. Each task in the multi-task network is responsible for brain parcellation at one scale. In this way, with the constraint from large scales, the accuracy of brain

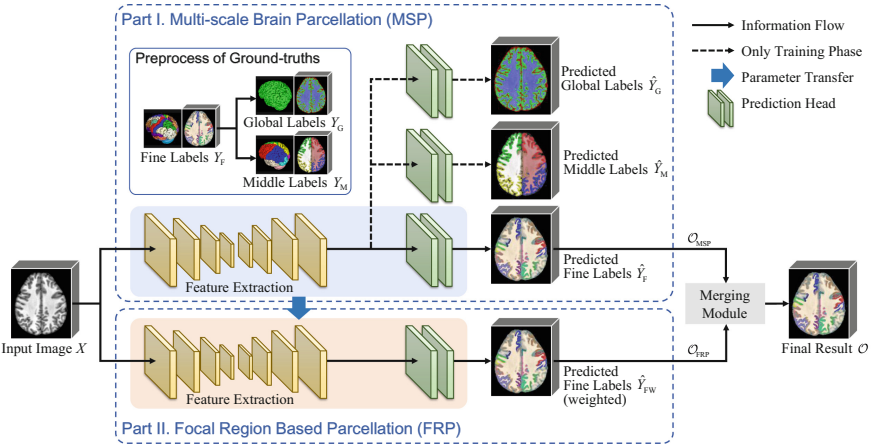


Fig. 1. The framework of our proposed multi-scale and focal region based deep learning network for fine brain parcellation.

parcellation at fine scale can be improved. Second, a focal region based auxiliary network is presented to further enhance the parcellation performance at the fine scale. In particular, brain regions, which are difficult to be parcellated, are given large weights in the loss function such that the auxiliary network can focus on these difficult brain regions. The final parcellation results are produced by a merging module using the outputs from the multi-task and auxiliary networks. The main contributions of this work are summarized as follows:

- We propose a novel multi-scale brain parcellation network (MSP) to parcellate brain regions at global, middle, and fine scales simultaneously, by which the performance of fine brain parcellation can be enhanced with high computational efficiency.
- We present a focal region based parcellation (FRP) network to boost the performance of brain regions that are difficult to be parcellated.
- We introduce a merging module (MM) to fuse the results of MSP and FRP networks to get accurate and robust brain parcellation results.

2 Method

The overview of our proposed multi-scale and focal region based parcellation framework is shown in Fig. 1. It is composed of three main components: 1) multi-scale brain parcellation network, 2) focal region based parcellation network, and 3) merging module. Details of each component are described in the following sections.

2.1 Multi-scale Brain Parcellation Network

As aforementioned, fine brain parcellation involves a large number of brain regions, making accurate parcellation a difficult task. Actually, according to

anatomical knowledge, fine-scale brain regions can be merged into middle-scale brain regions, such as lobes or subcortical nuclei, which can be further merged into global-scale labels, such as white matter and gray matter. Therefore, instead of directly parcellating a large number of brain regions at fine scale, we propose to conduct the parcellation at multiple scales, such as middle and global scales.

As shown in Part I of Fig. 1, the multi-scale brain parcellation network is in a multi-task structure, where each task is responsible for brain parcellation at different scale, i.e., global, middle, or fine scale. All tasks share the same feature extraction network and have different prediction heads (see Fig. 1). In this way, with the constraints from parcellation among different scales, features in the feature extraction network are learned under the constraints of hierarchical relations between different scales, making the brain region related features more discriminative and robust as compared with the features learned for one-scale brain parcellation. The loss function of the multi-scale brain parcellation (MSP) network is defined as:

$$\mathcal{L}_{\mathrm{MSP}} = \alpha_1 \mathcal{L}(\hat{Y}_{\mathrm{G}}, Y_{\mathrm{G}}) + \alpha_2 \mathcal{L}(\hat{Y}_{\mathrm{M}}, Y_{\mathrm{M}}) + \alpha_3 \mathcal{L}(\hat{Y}_{\mathrm{F}}, Y_{\mathrm{F}}), \tag{1}$$

where $\hat{Y}_{\mathrm{G}}$, $\hat{Y}_{\mathrm{M}}$, $\hat{Y}_{\mathrm{F}}$ represent parcellation results at global, middle and fine scales, Y_{G}, Y_{M}, and Y_{F} represent their corresponding ground-truth parcellations. $\alpha_i\,(i = 1, 2, 3)$ are weight parameters of three scales. Cross-entropy loss is adopted as $\mathcal{L}(.)$. The weight parameters are set empirically according to the relative difficulty and importance of each task. Here, $\alpha_1, \alpha_2, \alpha_3$ are set to 0.1, 0.2, and 0.7, respectively.

2.2 Focal Region Based Parcellation

Since some brain regions often have unclear anatomical boundaries or their volumes are too small, such brain regions are difficult to be detected at fine scale. To this end, we propose an auxiliary focal region based parcellation (FRP) network (see Part II of Fig. 1) to focus on the brain regions difficult to be parcellated at fine scale.

Specifically, we first train the MSP network in part I and use it in the inference phase to get the parcellation results of validation set at fine scale, based on which the mean Dice score of each brain region can be calculated. According to the mean Dice scores, brain regions which have small mean Dice scores are regarded as difficult regions to be parcellated, and large weights are assigned to these brain regions in the loss function of the FRP network. In this way, the performance of parcellation in these brain regions could be boosted. Specifically, the loss function of the FRP network is defined as:

$$\mathcal{L}_{\mathrm{FRP}} = \sum_{k=0}^{K} w_k \times l_k, \tag{2}$$

where K is the number of brain regions, l_k is the loss of the k-th brain region, w_k is the corresponding weight assigned to it, and l_k is the cross-entropy loss

between the parcellation result and the ground truth of the k-th brain region. The weight w_k is calculated by:

$$w_k = (1-\delta)(1-\frac{x_k - x_{\min}}{x_{\max} - x_{\min}}) + \delta \qquad (k = 0, 1, \cdots, K), \tag{3}$$

where $w_k \in [\delta, 1]$, $0 \leq \delta < 1$. $x_{\min}$ and $x_{\max}$ are the smallest and biggest Dice scores of all regions, respectively. x_k is the Dice score of the k-th brain region, which reflects the difficulty of labeling this region. The smaller of x_k, the harder of the k-th brain region, thus more attention should be paid to it, i.e., using larger w_k. δ is a hyper-parameter used to control relationship between the normalized x_k and weight w_k. The smaller of δ, the bigger range of w_k. In the experiment, δ is set to 0.2 for both stability and discrimination.

It is worth noting that, since the MSP network in Part I provides information about the difficulty of labeling brain regions that are needed by the FRP network in Part II, the MSP network is trained first. Furthermore, to provide a good initialization for the latter network, the network parameters in the former trained network (i.e., feature extraction and prediction head of fine-scale labels in MSP, indicated as the blue arrow in Fig. 1) are transferred to the latter network. The MSP network is fixed during the training of the whole framework, i.e., MSP+FRP, i.e., parameters of the FRP network are updated only.

2.3 Merging Module

As aforementioned, in our framework, the MSP network is used to get brain parcellation at fine scale, while the FRP network is responsible for producing brain parcellation at difficult brain regions. To get the final brain parcellation, a merging module (MM) is applied to combine the parcellation results from both MSP and FRP networks. Specifically, for an input brain image, the outputs of the MSP and FRP networks are $\mathcal{O}_{\text{MSP}}$ and $\mathcal{O}_{\text{FRP}}$, respectively. The final brain parcellation produced by the merging module is defined as:

$$\mathcal{O}(x) = \arg\max_k \Big(\max\big(\mathcal{O}_{\text{MSP}}(x,k), \mathcal{O}_{\text{FRP}}(x,k)\big)\Big), \tag{4}$$

where $\mathcal{O}_{\text{MSP}}(x,k)$ is the probability (after softmax) of voxel x belonging to the k-th brain region. $\mathcal{O}_{\text{FRP}}(x,k)$ is defined in the same way.

Based on the features learned by the MSP network with highly discriminative information of brain regions, as well as enhanced performance in difficult brain regions using the FRP network, high accurate final parcellation results can be produced after the merging module.

3 Experiments and Results

3.1 Dataset and Implementation Details

Dataset. A public multi-center dataset CoRR [17] containing T1 brain MR images of 267 subjects is used in our experiment. All the images go through

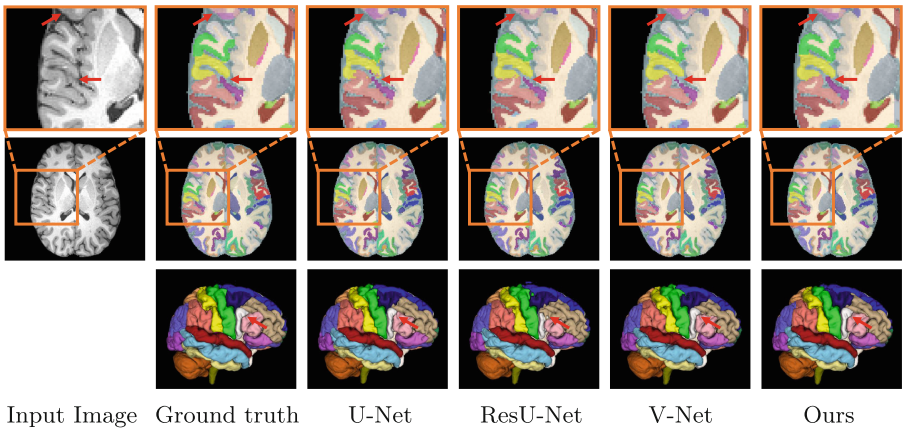


Fig. 2. Typical parcellation results using different methods under comparison. Bottom: 3D views of the parcellated brain. Middle: Axial views of the parcellated brain. Top: Zoomed views.

the preprocessing pipeline including skull stripping, bias correction, resampling to $1 \times 1 \times 1\,\mathrm{mm}^3$, and cropping or padding to the size of $256 \times 256 \times 256$. To create ground-truth of brain regions at fine scale, each image is labeled to 101 brain regions using FreeSurfer [4] and further revised by an experienced rater according to the DK atlas [3]. For each image, the labeled 101 brain regions are fused into 26 brain regions at middle scale and 3 brain regions at global scale according to the hierarchical anatomical definition of brain regions (as shown in Fig. 1).

Implementation Details. From the dataset, 200, 35, and 32 images are randomly selected and used for training, validation, and testing, respectively. Image patches of size $128 \times 128 \times 32$ are randomly cropped from images as input of our framework. Adam [7] is adopted as the optimization method with an initial learning rate of 0.0001 and drops 0.5x after 30, 50, and 70 epochs. The batch size and the maximal epoch number are set to 10 and 150, respectively. All experiments are implemented on PyTorch 1.10.0 and trained on one NVIDIA A100 GPU with a memory of 30 GB.

3.2 Comparison with State-of-the-Art Methods

Besides the proposed framework, three state-of-the-art methods, including U-Net [9], ResU-Net [16], and V-Net [8] are also evaluated in our experiment. Dice similarity coefficient (DSC) is used to quantitatively evaluate the parcellation accuracy.

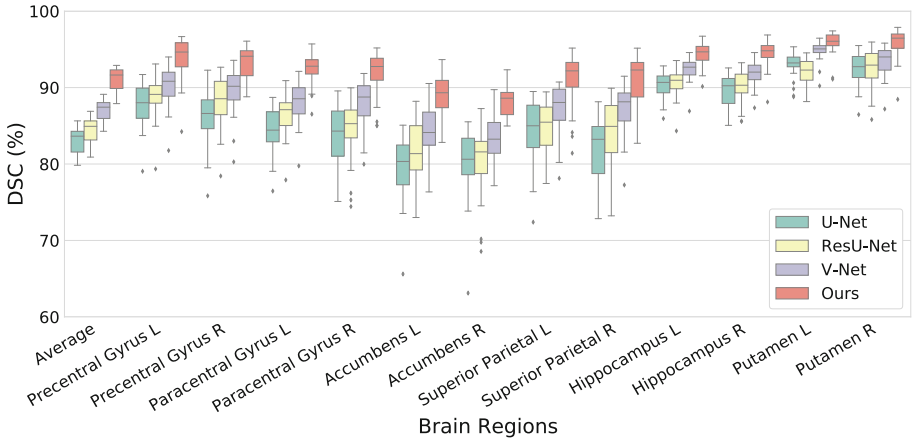


Fig. 3. Box plots of DSC (%) on average of 101 brain regions and also on 12 typical regions of different methods. "L" and "R" mean left and right, respectively. More results on all the regions can be found in Supplementary Materials.

Examples of parcellation results using each method are shown in Fig. 2. Obviously, the parcellation result of our framework is more accurate at both the brain region boundaries as well as small and difficult-to-parcellate regions.

Quantitative results are shown in Table 1, where mean and standard deviation of average DSC of 101 brain regions and DSC of 3 typical brain regions (left precentral gyrus, left hippocampus, and left accumbens) are calculated. For the average DSC of 101 brain regions, our approach outperformed U-Net, ResU-Net, and V-Net by 7.93%, 6.56%, and 4.07%, respectively. On cortical gyri like left precentral gyrus, or subcortical nuclei like left hippocampus and left accumbens, our method also outperformed the competing methods. Moreover, paired t-test is performed on the results of the average DSC of 101 brain regions, and the result shows that the improvement by our method is statistically significant (p-values < 0.01). Box plots of DSC on average of 101 brain regions and DSC of 12 typical brain regions are shown in Fig. 3. Clearly, our approach outperforms the other three methods. Particularly, for the difficult regions with relatively low DSC, our approach gets more improvement in accuracy. Due to the space limit, more results can be found in Supplementary Materials.

3.3 Ablation Study

Ablation studies are conducted to evaluate the effectiveness of the three main components proposed in our framework, i.e., MSP, FRP and MM. Quantitative results are shown in Table 2. The baseline network is 3D V-Net, which outputs brain parcellation results at fine scale directly. The configurations and evaluation results of the ablation experiments can be found in the *module* column of Table 2. It can be seen that the multi-scale brain parcellation (MSP) network has significant improvement compared with the baseline network, *not only* on

Table 1. Comparison of results by our approach and three state-of-the-art methods. Mean ± standard deviation of DSC (%) on average of 101 brain regions and also on three typical regions are shown. Best results are in bold.

Method	Average	Left precentral gyrus	Left hippocampus	Left accumbens
U-Net	83.13 ± 1.70	87.64 ± 2.79	90.31 ± 1.74	79.59 ± 3.99
ResU-Net	84.50 ± 1.64	88.71 ± 2.48	90.57 ± 1.71	81.69 ± 4.04
V-Net	86.99 ± 1.45	90.36 ± 2.33	92.43 ± 1.39	84.56 ± 3.28
Ours	$\mathbf{91.06 \pm 1.60}$	$\mathbf{94.02 \pm 2.44}$	$\mathbf{94.36 \pm 1.47}$	$\mathbf{89.02 \pm 2.61}$

average of all regions *but also* on the other typical regions. By introducing the focal region based parcellation (FRP) network to baseline network + MSP, the performance can be further improved. The highest performance is achieved by integrating the outputs of MSP and FRP networks using the merging module (MM).

Table 2. Results of the ablation studies. Mean ± standard deviation of DSC (%) on average of 101 brain regions and also on three typical regions are reported. The baseline network is 3D V-Net that outputs fine-scale labels directly. "MSP", "FRP", and "MM" represent multi-scale brain parcellation network, focal region based parcellation network, and merging module, respectively. All three components play positive roles.

Module				Average	Left precentral gyrus	Left hippocampus	Left accumbens
Baseline	MSP	FRP	MM				
✓				86.99 ± 1.45	90.36 ± 2.33	92.43 ± 1.39	84.56 ± 3.28
✓	✓			90.24 ± 1.62	93.63 ± 2.48	93.99 ± 1.50	88.54 ± 2.78
✓	✓	✓		90.92 ± 1.62	93.95 ± 2.36	94.26 ± 1.54	88.54 ± 2.64
✓	✓	✓	✓	$\mathbf{91.06 \pm 1.60}$	$\mathbf{94.02 \pm 2.44}$	$\mathbf{94.36 \pm 1.47}$	$\mathbf{89.02 \pm 2.61}$

4 Conclusion

In this paper, we proposed a multi-scale and focal region based framework for fine brain parcellation. In particular, with constraints from parcellation at middle and global scales, the brain parcellation performance at fine scale (with large number of brain regions) can be effectively enhanced. Moreover, a novel focal region based parcellation network was proposed to force the network focus on brain regions difficult to be parcellated. Finally, a merging module was integrated to aggregate the outputs of the two networks and produce the final parcellation results of all fine brain regions. Experimental results demonstrated that the proposed framework outperforms the state-of-the-art methods under comparison with statistical significance.

Acknowledgement. This work was supported in part by National Key Research and Development Program of China (Grant No. 2017YFA0700800), National Natu-

ral Science Foundation of China (Grant No. 62131015, 62088102 and 62073012), Science and Technology Commission of Shanghai Municipality (STCSM) (Grant No. 21010502600), Natural Science Basic Research Plan in Shaanxi Province of China (Grant No. 2022JM-324), Key Research and Development Program of Shaanxi Province of China (Grant No. 2020GXLHY-008), Social Science Foundation of Shaanxi Province of China (Grant No. 2021K014), and Beijing Municipal Natural Science Foundation (Grant No. 7222307).

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
2. Coupé, P., et al.: AssemblyNet: a large ensemble of CNNs for 3D whole brain MRI segmentation. NeuroImage **219**, 117026 (2020)
3. Desikan, R.S., et al.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage **31**(3), 968–980 (2006)
4. Fischl, B.: FreeSurfer. NeuroImage **62**(2), 774–781 (2012)
5. Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M.: FastSurfer-a fast and accurate deep learning based neuroimaging pipeline. NeuroImage **219**, 117012 (2020)
6. Huo, Y., et al.: 3D whole brain segmentation using spatially localized atlas network tiles. NeuroImage **194**, 105–119 (2019)
7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
10. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N., et al.: QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage **186**, 713–727 (2019)
11. Roy, A.G., Conjeti, S., Sheet, D., Katouzian, A., Navab, N., Wachinger, C.: Error corrective boosting for learning fully convolutional networks with limited data. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 231–239. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_27
12. Sun, L., Shao, W., Wang, M., Zhang, D., Liu, M.: High-order feature learning for multi-atlas based label fusion: application to brain segmentation with MRI. IEEE Trans. Image Process. **29**, 2702–2713 (2020)
13. Sun, L., Shao, W., Zhang, D., Liu, M.: Anatomical attention guided deep networks for ROI segmentation of brain MR images. IEEE Trans. Med. Imaging **39**(6), 2000–2012 (2020)
14. Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A.: Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 611–623 (2013)

15. Wu, Z., Guo, Y., Park, S.H., Gao, Y., Dong, P., Lee, S.W., Shen, D.: Robust brain ROI segmentation by deformation regression and deformable shape model. Med. Image Anal. **43**, 198–213 (2018)
16. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted Res-UNet for high-quality retina vessel segmentation. In: 2018 9th International Conference on Information Technology in Medicine and Education (ITME), pp. 327–331. IEEE (2018)
17. Zuo, X.N., et al.: An open science resource for establishing reliability and reproducibility in functional connectomics. Sci. Data **1**(1), 1–13 (2014)

Author Index

Ahmad, Sahar 220
Ali, Sharib 101, 161
Alis, Deniz 425
Asad, Zuhayr 298
Azad, Reza 377

Bagci, Ulas 171
Bahrami, Mojtaba 387
Bai, Ling 111
Bartlett, Joseph 151
Berkovsky, Shlomo 348
Bhalerao, Abhir 249
Bhattarai, Binod 161
Bischof, Horst 328
Bonheur, Savinien 328
Boyer, Edmond 307
Burns, Andrew H. 269

Cansever, Aysel Türkvatan 425
Cao, Yifan 81
Carrión, Héctor 446
Che, Jifei 131
Chen, Chaoyue 91
Chen, Lizhou 230
Chen, Qingyu 11
Chen, Xi 111
Chen, Yongwei 239
Chen, Yuntian 230
Cheng, Jiale 406
Cheng, Jun 210
Cheng, Pujin 32
Chew, Emily Y. 11
Chung, Albert C. S. 72
Chow, Nikki Ka Wai 210
Collins, Toby 121
Cong, Cong 348

Dai, Haixing 269, 279
Dan, Ruilong 81
Dang, Kang 239
Daniel Oldan, Jorge 358
Deng, Ruining 298
Di Ieva, Antonio 348
Ding, Xiaowei 239
Dong, Pei 131
Dorent, Reuben 288
Du, Lei 191
Du, Shaoyi 466
Duan, Jinming 151

East, James E. 101

Fang, Yuqi 358
Feng, Dehua 111
Fogo, Agnes B. 298

Gajowczyk, Milosz 318
Ge, Yuyan 466
Ghorbani, Mahsa 387
Gilmore, John H. 406
Glicksberg, Benjamin S. 11
Gomez, Marcella 446
Gorden, Mae 436
Gorthi, Subrahmanyam 397
Graham, Simon 249
Gray Yarbrough, Wendell 358
Grodek, Piotr 318
Gu, Yun 62
Guan, Hao 201
Güler, Arda 425
Gupta, Soumya 101, 161

He, Huaqing 32
He, Junjun 62
He, Zhibin 191
Hebbar, Akarsh 269
Heidari, Moein 377
Hostettler, Alexandre 121
Huang, Weimin 210
Huang, Yongsong 43
Huo, Yuankai 298

Isaeva, Natalia 358
Isseroff, Roslyn Rivkah 446

Jafari, Mohammad 446
Jahanifar, Mostafa 249
Jeevan, Govind 140

Jewsbury, Robert 249
Jia, Gangyong 81
Jia, Xi 151
Jiang, Caiwen 466
Jiang, Xiajun 338
Jin, Ze 456
Joutard, Samuel 288

Kang, Yanqing 191
Karaarslan, Ercan 425
Kass, Michael 436
Keenan, Tiarnan D. L. 11
Konopczynski, Tomasz 318
Korbecki, Adrian 318

Lankireddy, Prabhat 397
Lee, Junghwan 11
Leedham, Simon 101
Lei, Chenjin 131
Li, Gang 220, 406
Li, Qing 279
Li, Quanzheng 269
Li, Sheng 269
Li, Xiang 269
Li, Yunxiang 81
Li, Zheren 131
Lin, Li 32
Lin, Mingquan 436
Lin, Weili 201, 220, 358, 406
Lin, Zhiping 210
Liu, Jingshuai 21
Liu, Lei 436
Liu, Mingxia 1, 201, 259, 358
Liu, Sidong 348
Liu, Siyuan 201
Liu, Tianming 269, 279
Liu, Wei 269
Liu, Zhengliang 269, 279
López Diez, Paula 416
Lu, Wenqi 151
Lu, Zhiyong 11
Luo, Xiangde 81
Lv, Jiahuan 111

Ma, Lei 466
Margeta, Jan 416
Meng, Di 307
Merhof, Dorit 377
Minhas, Fayyaz 249
Missel, Ryan 338
Modat, Marc 288
Mohammed, Eslam 307
Mok, Tony C. W. 72
Moradi, Mehdi 53
Mou, Xuanqin 111

Nan, Fang 220
Navab, Nassir 387
Negahdar, Mohammadreza 181
Nguyen, Ethan H. 298
Norouzi, Narges 446

Öksüz, İlkay 425
Olschewski, Horst 328
Omachi, Shinichiro 43
Otani, Niels 338
Ourselin, Sebastien 288
Özer, Caner 425

Pagnucco, Maurice 348
Pan, Liming 279
Parke Schrank, Travis 358
Patou, François 416
Paulsen, Rasmus R. 416
Pawan, S. J. 140
Peng, Linkai 32
Peng, Yifan 11, 436
Pienn, Michael 328
Podgorski, Przemyslaw 318
Potter, Guy G. 259
Pujades, Sergi 307

Qiao, Lishan 1
Qin, Chen 21
Qiu, Zhaowen 151

Radutoiu, Ana-Teodora 416
Rajan, Jeny 140
Rajpoot, Nasir 249
Rambour, Clément 121
Raza, Shan E. Ahmed 249
Rezayi, Saed 269
Rittscher, Jens 101, 161
Rolandi, Marco 446
Rygiel, Patryk 318

Shen, Dinggang 131, 230, 466
Shi, Cheng 279
Shi, Enze 191
Shi, Feng 230, 466

Shu, Xin 91
Sindhura, Chitimireddy 397
Sobanski, Michal 318
Song, Bin 230
Song, Yang 348
Steffens, David C. 259
Sun, Hui 62
Suzuki, Kenji 456

Tan, Chenghao 81
Tan, Cher Heng 210
Tan, Hongna 131
Tang, Xiaoying 32
Tang, Zhenyu 466
Tassel, Sarah Van 436
Themyr, Loic 121
Thome, Nicolas 121
Toloubidokhti, Maryam 338
Turney, Ben 161

Urschler, Martin 328

Vercauteren, Tom 288
Vu, Quoc Dang 249

Wang, Fei 11, 436
Wang, Li 81, 220, 406
Wang, Lihong 259
Wang, Linwei 338
Wang, Meiyun 131
Wang, Qian 230
Wang, Qianqian 1
Wang, Qingzhong 43
Wang, Ruoyang 191
Wang, Shuai 81
Wang, Xiaohong 210
Wang, Yaqi 81
Wang, Zizhou 91
Wanyan, Tingyi 11
Wong, Ken C. L. 53
Wu, Di 220
Wu, Jinru 191
Wu, Qingxia 131
Wu, Rong 259
Wu, Xia 279
Wu, Yaping 131
Wu, Ye 220
Wu, Yuli 377
Wu, Zhengwang 220, 406
Wu, Zihao 269, 279

Xia, Yong 367
Xu, Jianguo 91
Xu, Ziang 101, 161
Xue, Zhong 131, 230

Yaghoobi, Mehrdad 21
Yang, Haichun 298
Yang, Hsin-Ya 446
Yang, Jie 62
Yang, Yixing 348
Yang, Yuqiao 456
Yao, Linlin 230
Yap, Pew-Thian 201, 220
Yeganeh, Yousef 387
You, Xin 62
Yu, Minhui 259
Yu, Qinji 239
Yu, Sigang 191
Yu, Yijie 91
Yuan, Xinrui 406
Yuen, Peter Ho Hin 210

Zhan, Yiqiang 230
Zhang, Han 72
Zhang, Jianpeng 367
Zhang, Lei 91
Zhang, Lintao 259
Zhang, Lu 279
Zhang, Shaoteng 367
Zhang, Shu 111, 191
Zhang, Tianyang 151
Zhang, Tuo 191
Zhang, Xin 406
Zhang, You 81
Zhang, Zheyuan 171
Zhao, Fenqiang 406
Zhao, Lin 269, 279
Zhao, Shijie 279
Zhou, Huiyu 81
Zhou, Zhiguo 111
Zhou, Ziyu 239
Zhu, Dajiang 269, 279