

# **International Conference on Medical Imaging with Deep Learning**

**Volume 102**



*M. Jorge Cardoso, Aasa Feragen, Ben Glocker, Ender Konukoglu, Ipek*

*Oguz, Gozde Unal and Tom Vercauteren*

## Table of Contents

<b>Preface</b>	1
<i>AnatomyGen: Deep Anatomy Generation From Dense Representation With Applications in Mandible Synthesis</i>	4
A. Abdi, H. Borgard, P. Abolmaesumi & S. Fels; PMLR 102:4–14, 2019.	
<i>Exploring local rotation invariance in 3D CNNs with steerable filters</i>	15
V. Andrearczyk, J. Fageot, V. Oreiller, X. Montet & A. Depeursinge; PMLR 102:15–26, 2019.	
<i>On the Spatial and Temporal Influence for the Reconstruction of Magnetic Resonance Fingerprinting</i>	27
F. Balsiger, O. Scheidegger, P.G. Carlier, B. Marty & M. Reyes; PMLR 102:27–38, 2019.	
<i>Image Synthesis with a Convolutional Capsule Generative Adversarial Network</i>	39
C. Bass, T. Dai, B. Billot, K. Arulkumaran, A. Creswell, C. Clopath, V. De Paola & A.A. Bharath; PMLR 102:39–62, 2019.	
<i>Fusing Unsupervised and Supervised Deep Learning for White Matter Lesion Segmentation</i>	63
C. Baur, B. Wiestler, S. Albarqouni & N. Navab; PMLR 102:63–72, 2019.	
<i>Learning interpretable multi-modal features for alignment with supervised iterative descent</i>	73
M. Blendowski & M.P. Heinrich; PMLR 102:73–83, 2019.	
<i>Learning from sparsely annotated data for semantic segmentation in histopathology images</i>	84
J.-M. Bokhorst, H. Pinckaers, P. van Zwam, I. Nagtegaal, J. van der Laak & F. Ciompi; PMLR 102:84–91, 2019.	
<i>Segmenting Potentially Cancerous Areas in Prostate Biopsies using Semi-Automatically Annotated Data</i>	92
N. Burlutskiy, N. Pinchaud, F. Gu, D. Hägg, M. Andersson, L. Björk, K. Eurén, C. Svensson, L.K. Wilén & M. Hedlund; PMLR 102:92–108, 2019.	
<i>Deep Hierarchical Multi-label Classification of Chest X-ray Images</i>	109
H. Chen, S. Miao, D. Xu, G.D. Hager & A.P. Harrison; PMLR 102:109–120, 2019.	
<i>Digital Stained Confocal Microscopy through Deep Learning</i>	121
M. Combalia, J. Pérez-Anker, A. García-Herrera, L. Alos, V. Vilaplana, F. Marqués, S. Puig & J. Malvehy; PMLR 102:121–129, 2019.	

<i>Deep Reinforcement Learning for Subpixel Neural Tracking</i>	130
T. Dai, M. Dubois, K. Arulkumaran, J. Campbell, C. Bass, B. Billot, F. Uslu, V. de Paola, C. Clopath & A. Bharath; PMLR 102: <a href="#">130–150</a> , 2019.	
<i>Stain-Transforming Cycle-Consistent Generative Adversarial Networks for Improved Segmentation of Renal Histopathology</i>	151
T. de Bel, M. Hermsen, J. Kers, J. van der Laak & G. Litjens; PMLR 102: <a href="#">151–163</a> , 2019.	
<i>Learning joint lesion and tissue segmentation from task-specific hetero-modal datasets</i>	164
R. Dorent, W. Li, J. Ekanayake, S. Ourselin & T. Vercauteren; PMLR 102: <a href="#">164–174</a> , 2019.	
<i>Unsupervisedly Training GANs for Segmenting Digital Pathology with Automatically Generated Annotations</i>	175
M. Gadermayr, L. Gupta, B.M. Klinkhammer, P. Boor & D. Merhof; PMLR 102: <a href="#">175–184</a> , 2019.	
<i>Transfer Learning by Adaptive Merging of Multiple Models</i>	185
R. Geyer, L. Corinzia & V. Wegmayr; PMLR 102: <a href="#">185–196</a> , 2019.	
<i>Assessing Knee OA Severity with CNN attention-based end-to-end architectures</i>	197
M. Górriz, J. Antony, K. McGuinness, X. Giró-i-Nieto & N. O'Connor; PMLR 102: <a href="#">197–214</a> , 2019.	
<i>Iterative learning to make the most of unlabeled and quickly obtained labeled data in histology</i>	215
L. Gupta, B. Mara Klinkhammer, P. Boor, D. Merhof & M. Gadermayr; PMLR 102: <a href="#">215–224</a> , 2019.	
<i>Generative Image Translation for Data Augmentation of Bone Lesion Pathology</i>	225
A. Gupta, S. Venkatesh, S. Chopra & C. Ledig; PMLR 102: <a href="#">225–235</a> , 2019.	
<i>Cluster Analysis in Latent Space: Identifying Personalized Aortic Valve Prosthesis Shapes using Deep Representations</i>	236
J. Hagenah, K. Kühl, M. Scharfschwerdt & F. Ernst; PMLR 102: <a href="#">236–249</a> , 2019.	
<i>Sparse Structured Prediction for Semantic Edge Detection in Medical Images</i>	250
L. Hansen & M. Heinrich; PMLR 102: <a href="#">250–259</a> , 2019.	
<i>Exclusive Independent Probability Estimation using Deep 3D Fully Convolutional DenseNets: Application to IsoIntense Infant Brain MRI Segmentation</i>	260
S.R. Hashemi, S.P. Prabhu, S.K. Warfield & A. Gholipour; PMLR 102: <a href="#">260–274</a> , 2019.	
<i>Dynamic MRI Reconstruction with Motion-Guided Network</i>	275
Q. Huang, D. Yang, H. Qu, J. Yi, P. Wu & D. Metaxas; PMLR 102: <a href="#">275–284</a> , 2019.	
<i>Boundary loss for highly unbalanced segmentation</i>	285
H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz & I. Ben Ayed; PMLR 102: <a href="#">285–296</a> , 2019.	

<i>Neural Processes Mixed-Effect Models for Deep Normative Modeling of Clinical Neuroimaging Data</i>	297
S. Kia & A. Marquand; PMLR 102:297–314, 2019.	
<i>Capturing Single-Cell Phenotypic Variation via Unsupervised Representation Learning</i>	315
M.W. Lafarge, J.C. Caicedo, A.E. Carpenter, J.P. Pluim, S. Singh & M. Veta; PMLR 102:315–325, 2019.	
<i>DavinciGAN: Unpaired Surgical Instrument Translation for Data Augmentation</i>	326
K. Lee, M.-K. Choi & H. Jung; PMLR 102:326–336, 2019.	
<i>Dense Segmentation in Selected Dimensions: Application to Retinal Optical Coherence Tomography</i>	337
B. Liefers, C. González-Gonzalo, C. Klaver, B. van Ginneken & C. Sánchez; PMLR 102:337–346, 2019.	
<i>Dynamic Pacemaker Artifact Removal (DyPAR) from CT Data using CNNs</i>	347
T. Lossau (née Elss), H. Nickisch, T. Wissel, S. Hakmi, C. Spink, M. Morlock & M. Grass; PMLR 102:347–357, 2019.	
<i>Group-Attention Single-Shot Detector (GA-SSD): Finding Pulmonary Nodules in Large-Scale CT Images</i>	358
J. Ma, X. Li, H. Li, B.H. Menze, S. Liang, R. Zhang & W.-S. Zheng; PMLR 102:358–369, 2019.	
<i>A novel segmentation framework for uveal melanoma in magnetic resonance imaging based on class activation maps</i>	370
H.-G. Nguyen, A. Pica, J. Hrbacek, D. Weber, F.L. Rosa, A. Schalenbourg, R. Sznitman & M. Bach Cuadra; PMLR 102:370–379, 2019.	
<i>High-quality segmentation of low quality cardiac MR images using k-space artefact correction</i>	380
I. Oksuz, J. Clough, W. Bai, B. Ruijsink, E. Puyol-Antón, G. Cruz, C. Prieto, A.P. King & J.A. Schnabel; PMLR 102:380–389, 2019.	
<i>Weakly Supervised Deep Nuclei Segmentation using Points Annotation in Histopathology Images</i>	390
H. Qu, P. Wu, Q. Huang, J. Yi, G.M. Riedlinger, S. De & D.N. Metaxas; PMLR 102:390–400, 2019.	
<i>Joint Learning of Brain Lesion and Anatomy Segmentation from Heterogeneous Datasets</i>	401
N. Roulet, D.F. Slezak & E. Ferrante; PMLR 102:401–413, 2019.	
<i>Learning with Multitask Adversaries using Weakly Labelled Data for Semantic Segmentation in Retinal Images</i>	414
O. Saha, R. Sathish & D. Sheet; PMLR 102:414–426, 2019.	
<i>MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty</i>	427
R. Shaw, C. Sudre, S. Ourselin & M.J. Cardoso; PMLR 102:427–436, 2019.	

TABLE OF CONTENTS

<i>A Hybrid, Dual Domain, Cascade of Convolutional Neural Networks for Magnetic Resonance Image Reconstruction</i>	437
R. Souza, R.M. Lebel & R. Frayne; PMLR 102: <a href="#">437–446</a> , 2019.	
<i>3D multirater RCNN for multimodal multiclass detection and characterisation of extremely small objects</i>	447
C. Sudre <i>et al</i> ; PMLR 102: <a href="#">447–456</a> , 2019.	
<i>XLSor: A Robust and Accurate Lung Segmentor on Chest X-Rays Using Criss-Cross Attention and Customized Radiorealistic Abnormalities Generation</i>	457
Y.-B. Tang, Y.-X. Tang, J. Xiao & R.M. Summers; PMLR 102: <a href="#">457–467</a> , 2019.	
<i>Training Deep Networks on Domain Randomized Synthetic X-ray Data for Cardiac Interventions</i>	468
D. Toth, S. Cimen, P. Ceccaldi, T. Kurzendorfer, K. Rhode & P. Mountney; PMLR 102: <a href="#">468–482</a> , 2019.	
<i>Prediction of Disease Progression in Multiple Sclerosis Patients using Deep Learning Analysis of MRI Data</i>	483
A. Tousignant, P. Lemaître, D. Precup, D.L. Arnold & T. Arbel; PMLR 102: <a href="#">483–492</a> , 2019.	
<i>Learning beamforming in ultrasound imaging</i>	493
S. Vedula, O. Senouf, G. Zurakhov, A. Bronstein, O. Michailovich & M. Zibulevsky; PMLR 102: <a href="#">493–511</a> , 2019.	
<i>Adversarial Pseudo Healthy Synthesis Needs Pathology Factorization</i>	512
T. Xia, A. Chartsias & S.A. Tsaftaris; PMLR 102: <a href="#">512–526</a> , 2019.	
<i>VOCA: Cell Nuclei Detection In Histopathology Images By Vector Oriented Confidence Accumulation</i>	527
C. Xie, C.M. Vanderbilt, A. Grabenstetter & T.J. Fuchs; PMLR 102: <a href="#">527–539</a> , 2019.	
<i>Unsupervised Lesion Detection via Image Restoration with a Normative Prior</i>	540
S. You, K.C. Tezcan, X. Chen & E. Konukoglu; PMLR 102: <a href="#">540–556</a> , 2019.	
<i>Deep Learning Approach to Semantic Segmentation in 3D Point Cloud Intra-oral Scans of Teeth</i>	557
F. Ghazvinian Zanjani, D. Anssari Moin, B. Verheij, F. Claessen, T. Cherici, T. Tan & P. de With; PMLR 102: <a href="#">557–571</a> , 2019.	
<i>SPDA: Superpixel-based Data Augmentation for Biomedical Image Segmentation</i>	572
Y. Zhang, L. Yang, H. Zheng, P. Liang, C. Mangold, R.G. Loreto, D.P. Hughes & D.Z. Chen; PMLR 102: <a href="#">572–587</a> , 2019.	
<i>CARE: Class Attention to Regions of Lesion for Classification on Imbalanced Data</i>	588
J. Zhuang, J. Cai, R. Wang, J. Zhang & W. Zheng; PMLR 102: <a href="#">588–597</a> , 2019.	

## Preface

This volume contains the Proceedings of the Second International Conference on Medical Imaging with Deep Learning – MIDL 2019. The conference was organized jointly by Imperial College London, UK and King’s College London, UK, with a team of program chairs from University of Copenhagen/Technical University of Denmark, Denmark, ETH Zürich, Switzerland, Istanbul Technical University, Turkey, and Vanderbilt University, USA. The conference was held between July 8 and 10, 2019, at Imperial College’s South Kensington Campus in London, UK.

MIDL 2019 attracted word-class researchers, scientists, engineers, as well as clinicians, who are keen on developing novel algorithms to solve medical imaging problems using deep learning techniques. The conference had two submission tracks: *full paper* and *extended abstract*. Through the OpenReview system, 117 full paper and 158 extended abstract submissions were received. For both tracks, the review process was single blind by default but the authors were also given the option to keep their identities anonymous. All submitted papers were on display at the OpenReview system from the time of their submission. All reviews and rebuttals were made publicly available after the decision process. Authors were however given the option of removing rejected submissions from the system.

For the full paper track, the review process was initiated by the Program Chairs (PC), and was handled by one primary Area Chair (AC) for each paper. The AC assigned exactly three expert reviewers, whose identities were kept anonymous. Area chairs wrote an additional report based on the submission, the anonymous reviews and the rebuttals. Submissions that were proposed to be accepted by all the reviewers and area chairs were accepted without further discussion by the PC. There were 31 such *early accepted* submissions (26.5% of all submissions). Likewise, submissions that were proposed to be rejected by all the reviewers and area chairs were rejected without further discussion by the PC. There were 33 such *early rejected* submissions (28.2% of all submissions). The remaining 53 articles were further discussed in a PC meeting that included the program chairs as well as the conference chairs. As a result of this meeting, out of these 53 discussed papers, 16 submissions were accepted and 37 were rejected. The decisions were based on careful examination of the submission, AC and reviewer comments and the rebuttals. The final acceptance rate of the full paper tract was 40.2%. Among the 47 accepted full paper articles, 19 were selected as oral presentations (16.2% of all submissions), based on reviewer and AC proposals, and 28 for poster presentations.

Acknowledging the benefits of sharing novel ideas at their infancy, at MIDL 2019, extended abstract submissions were reviewed with a more lightweight process by exactly two reviewers, whose identities were kept anonymous. The focus of the review was put on the potential for encouraging constructive and thought-provoking discussions at the conference. The acceptance rate for the extended abstract track was kept higher than for the full papers to encourage a wide representation of the field. All submissions proposed to be accepted based on the average score of the reviewers weighted by reviewer confidence were thus accepted. Among the 158 submitted abstracts, 105 abstracts were accepted to be presented as posters at the conference (66.5% acceptance rate). These

proceedings only contain the accepted full paper submissions, the extended abstracts can be found in the OpenReview system or in the corresponding arXiv compendium.

The articles in these proceedings are presented in alphabetical order by first author name. The papers comprise of a wide range of topics including new deep learning architectures, loss functions, geometric learning, domain adaptation, transfer learning, and applications to lesion segmentation, brain imaging, cardiac imaging, neurodegenerative diseases and image reconstruction.

The PC and organizing chairs would like to thank the OpenReview staff for their support in hosting the submission site and PMLR staff for their support in finalizing these proceedings.

Most importantly, we would like to thank our Area Chairs, and Reviewers for their hard work in preparing and helping us shape the final technical program of MIDL 2019. The conference and the current proceedings are a result of their work.

Finally, we would like to thank all our sponsors for the financial support, which made the MIDL 2019 conference possible. We look forward to seeing you in Montreal, Canada, in 2020 for the next edition of MIDL!

*May 23, 2019*

The Editorial Team:

M. Jorge Cardoso  
 King's College London, UK  
[m.jorge.cardoso@kcl.ac.uk](mailto:m.jorge.cardoso@kcl.ac.uk)

Aasa Feragen  
 Technical University of Denmark, Denmark  
 University of Copenhagen, Denmark  
[aasa@diku.dk](mailto:aasa@diku.dk)

Ben Glocker  
 Imperial College London, UK  
[b.glocker@imperial.ac.uk](mailto:b.glocker@imperial.ac.uk)

Ender Konukoglu  
 ETH Zurich, Switzerland  
[ender.konukoglu@vision.ee.ethz.ch](mailto:ender.konukoglu@vision.ee.ethz.ch)

Ipek Oguz  
 Vanderbilt University, USA  
[ipek.oguz@vanderbilt.edu](mailto:ipek.oguz@vanderbilt.edu)

Gozde Unal  
 Istanbul Technical University, Turkey  
[gozde.unal@itu.edu.tr](mailto:gozde.unal@itu.edu.tr)

Tom Vercauteren  
 King's College London, UK  
[tom.vercauteren@kcl.ac.uk](mailto:tom.vercauteren@kcl.ac.uk)

**Keynote Chair** William M. Wells, Harvard Medical School

**Area Chairs** Ehsan Adeli, Suyash Awate, Arrate Munoz Barrutia, Christian Baumgartner, Ismail Ben Ayed, Hrvoje Bogunovic, Tessa Cook, Marleen de Bruijne, Qi Dou, Nicha Dvornek, Shireen El-habian, Ilker Hacihaliloglu, Matthias Heinrich, Yi Hong, Eugenio Iglesias, Ivana Isgum, Konstantinos Kamnitsas, Shella Keilholz, Minjeong Kim, Sila Kurugol, Herve Lombaert, Christian Ledig, Geert Litjens, Lena Maier-Hein, Diana Mateus, Bjoern Menze, Brent Munsell, Ana Namburete, Marc Niethammer, Lauren O'Donnell, Ozan Oktay, Islem Rekik, Holger Roth, Clarisa Sanchez, Thomas Schulz, Bram van Ginneken, Archana Venkataraman, Guorong Wu, Xiahai Zhuang, Lilla Zollei

**Reviewers** Seyed-Ahmad Ahmadi, Shadi Albarqouni, Daniel Alexander, Sharib Ali, Ryan Amelon, Michela Antonelli, Shekoofeh Azizi, Ulas Bagci, Wenjia Bai, Sophia Bano, Mathilde Bateson, kayhan Batmanghelich, Erik Bekkers, Nesrine Bnouni, Christopher Bridge, Esther Bron, Aaron Carass, Daniel Castro, Joshua Cates, Juan Cerrolaza, Shikha Chaganti, Catie Chang, Pierre Chatelein, Cheng Chen, Hao Chen, Xiaoran Chen, Francesco Ciompi, Dana Cobzas, Joseph Cohen, Adrian Dalca, Estibaliz Gómez de Mariscal, Bob De Vos, Cem Deniz, Adrien Deppeursinge, Christian Desrosiers, Jose Dolz, Niharika Dsouza, Zach Eaton-Rosen, Bernhard Egger, Enzo Ferrante, Ahmed Fetit, Aina Frau-Pascual, Adrian Galdran, Mingchen Gao, Sarah Gerard, Ali Gholipour, Stamatia Giannarou, Polina Golland, German Gonzalez, Karthik Gopinath, Pedro Gordaliza, Ricardo Guerrero, Adam Harrison, Raein Hashemi, Yoonmi Hong, Yipeng Hu, henkjan huisman, Yuankai Huo, Sang Hyun Park, Hayato Itoh, Colin Jacobs, Avelino Javer, Shuman Jia, Yueming Jin, Amod Jog, Samuel Kadoury, Bernhard Kainz, Neerav Karani, Hoel Kervadec, Seyed Mostafa Kia, Jaeil Kim, Namkug Kim, G nther Klambauer, Simon Kohl, Arinbj rn Kolbeinsson, Panagiotis Korfiatis, Kivanc Kose, Wouter Kouw, Sofia Ira Ktena, Loic Le Folgoc, Hansang Lee, Nikolas Lessmann, Wenqi Li, Xiaoxiao Li, Mingxia Liu, Le Lu, Da Ma, Dwarikanath Mahapatra, Klaus Maier-Hein, Matteo Mancini, Karttikeya Mangalam, Rashindra Manniesing, Dorit Merhof, Fausto Milletari, Sara Moccia, Pim Moeskops, Keelin Murphy, Hannes Nickisch, Dong Nie, Julia Noothout, Hirohisa Oda, Ilkay Oksuz, John Onofrey, Jos  Orlando, Firat Ozdemir, Bartlomiej Papiez, Sarah Parisot, Nick Pawlowski, Loic Peter, Caroline Petitjean, Sergey Plis, Raphael Prevost, Jonas Richiardi, Nicola Rieke, Olaf Ronneberger, Mohammad Sabokrou, Michiel Schaap, Benoit Scherrer, Jo Schlemper, Julia Schnabel, Andreas Schuh, Philipp Seeb ck, Anjany Sekuboyina, Raghavendra Selvan, Harshita Sharma, Debdoott Sheet, Hoo-Chang Shin, Matthew Sinclair, Ayushi Sinha, Ziga Spiclin, Jackson Steinkamp, Danail Stoyanov, Martin Styner, Carole Sudre, Takaaki Sugino, Heung-II Suk, Raphael Sznitman, Christine Tanner, Ryutaro Tanno, David Tellez, Martin Urschler, Thomas van den Heuvel, Gijs van Tulder, Harini Veeraraghavan, Mitko Veta, Valery Vishnevskiy, Chenglong Wang, Xiaosong Wang, Ross Whitaker, Matthias Wilms, Jelmer Wolterink, Weidi Xie, Daguang Xu, Qiuping Xu, Ziyue Xu, Dong Yang, Guang Yang, Xin Yang, Yixuan Yuan, Fan Zhang, Han Zhang, Ling Zhang, Miaomiao Zhang, Can Zhao, Qingyu Zhao, Guoyan Zheng, Sihang Zhou, Yanning Zhou, Juntang Zhuang, Majd Zreik, Silas  rtting, Maria A Zuluaga

# AnatomyGen: Deep Anatomy Generation From Dense Representation With Applications in Mandible Synthesis

**Amir H. Abdi<sup>1</sup>**

AMIRABDI@ECE.UBC.CA

**Heather Borgard<sup>1</sup>**

HEATHER.BORGARD@UBC.CA

**Purang Abolmaesumi<sup>1</sup>**

PURANG@ECE.UBC.CA

**Sidney Fels<sup>1</sup>**

SSFELS@ECE.UBC.CA

<sup>1</sup> *Electrical and Computer Engineering Department, University of British Columbia, Canada*

## Abstract

This work is an effort in human anatomy synthesis using deep models. Here, we introduce a deterministic deep convolutional architecture to generate human anatomies represented as 3D binarized occupancy maps (voxel-grids). The shape generation process is constrained by the 3D coordinates of a small set of landmarks selected on the surface of the anatomy. The proposed learning framework is empirically tested on the mandible bone where it was able to reconstruct the anatomies from landmark coordinates with the average landmark-to-surface error of 1.42 mm. Moreover, the model was able to linearly interpolate in the  $\mathbb{Z}$ -space and smoothly morph a given 3D anatomy to another. The proposed approach can potentially be used in semi-automated segmentation with manual landmark selection as well as biomechanical modeling. Our main contribution is to demonstrate that deep convolutional architectures can generate high fidelity complex human anatomies from abstract representations.

**Keywords:** Deep generative model, 3D convolutional neural network, Shape generation, Geometric morphometrics, Shape interpolation

## 1. Introduction

Underlying dynamics of musculoskeletal systems are often studied by means of computational modeling. These models provide insights into system properties which are otherwise impossible or difficult to measure directly such as muscle fiber excitations, joint forces, and internal stresses (Hicks et al., 2015). Advancements in numerical modeling, computer graphics, and biomedical imaging, as well as increased interdisciplinary collaborations, have turned computational modeling into a blooming field of research (Andersen et al., 2017).

To study the biomechanics of a particular subject or patient, researchers create generic templates of anatomies from image-driven measurements or cadaver data. These templates are then morphed or rescaled to form subject-specific models that represent characteristics of the individual. However, this is a challenging and labour-intensive process, contingent on the availability of segmentations and the quality of imaging.

Statistical shape models (SSM) have been extensively investigated in the context of large-scale analysis of anatomical shapes and image segmentation (Zhang and Golland, 2016). This family of algorithms almost universally utilize some form of principal component analysis (PCA) and mixture models to represent variabilities in the population. While these methods have the potential to morph a statistical template to an unseen sample, they are limited or regularized by the principal

modes of variation and their respective variances. SSM approaches depend on the correspondence or coregistration of the training shapes as well as the quality of mapping to the unseen data.

With recent advances in deep learning architectures, some methods have been established in synthesizing photorealistic 2D images (Goodfellow et al., 2014; Rosca et al., 2017). With the extension of deep learning models to the realm of 3D geometries, inspiring efforts are being made to learn deterministic or probabilistic dense representations of 3D shapes (Girdhar et al., 2016; Smith and Meger, 2017; Liu et al., 2017; Tatarchenko et al., 2017) and synthesize unseen objects from learned distributions (Wu et al., 2016) or from 2D images (Choy et al., 2016). The closest work to our current approach is that of Brock et al., which uses a low-fidelity variational autoencoder for voxel-based shape modeling and a 3D convolutional classifier for object detection (Brock et al., 2016).

While synthesizing realistic shapes and images with deceptive overall perceptions is an intriguing objective in computer graphics and computer vision, the healthcare and biomedical engineering communities are more concerned about the clinical implications of the generated data. These concerns go beyond 3D modeling and are often raised in response to any generative model, including adversarial learning paradigms used to generate MR, CT (Wolterink et al., 2017), and PET data (Pan et al., 2018), or vision solutions for synthesizing realistic skin lesions (Baur et al., 2018). In other efforts, the image information is encoded in a dense latent space to regularize segmentation and super-resolution tasks (Oktay et al., 2018). Regardless, an association between the generated data and their clinical value is currently missing from the generative models in healthcare (Kazeminia et al., 2018).

To this end, we propose a deterministic generative convolutional approach to synthesize 3D anatomies conditioned on clinically relevant dense representations. The performance of this method is demonstrated on the human mandible bone because of its complexities due to thin plates, and multiple concavities and processes. The trained deep model is evaluated on its ability to generate new mandible samples from unseen landmark coordinates. Moreover, we demonstrate the model’s potential in smoothly interpolating in-between shapes in the dense space. The goal here is neither to morph a template average shape, nor to segment anatomy, but to map representations of the 3D shape from an abstract dense  $\mathbb{Z}$ -space to the 3D voxel-grid space.

## 2. Method

### 2.1. Network Architecture

The generative model proposed here, which we formally refer to as  $g(z)$ , is a deep neural network which deterministically maps a dense vectorized representation,  $z$ , to a cubic tensor,  $V$ . This function can be summarized as  $g : \mathbb{Z}^t \rightarrow \mathbb{V}$ , where  $\mathbb{V}$  represents the 3D binarized voxel-grid where voxels belonging to the object are assigned the value of 1.

The highest performing network architecture that we experimented with is depicted in Figure 1. The network starts with three dense layers, with no non-linear activations, which linearly combine the elements of the input dense vector of size  $|\mathbb{Z}|$  and form a bigger feature vector of size  $64 \times |\mathbb{Z}|$ , which is subsequently reshaped into a 4D tensor of shape  $|\mathbb{Z}| \times 4 \times 4 \times 4$ .

The dense layers are followed by several transposed strided 3D convolutions (TConv3D), which are generally referred to as deconvolutions and implemented as gradient of equivalent convolutional layers. Except for the first TConv3D layer with a kernel size of 1 and stride of 1, all other TConv3D layers have a kernel size of 4 and a stride of 2 across all the 3 axes. All layers are followed by Contin-

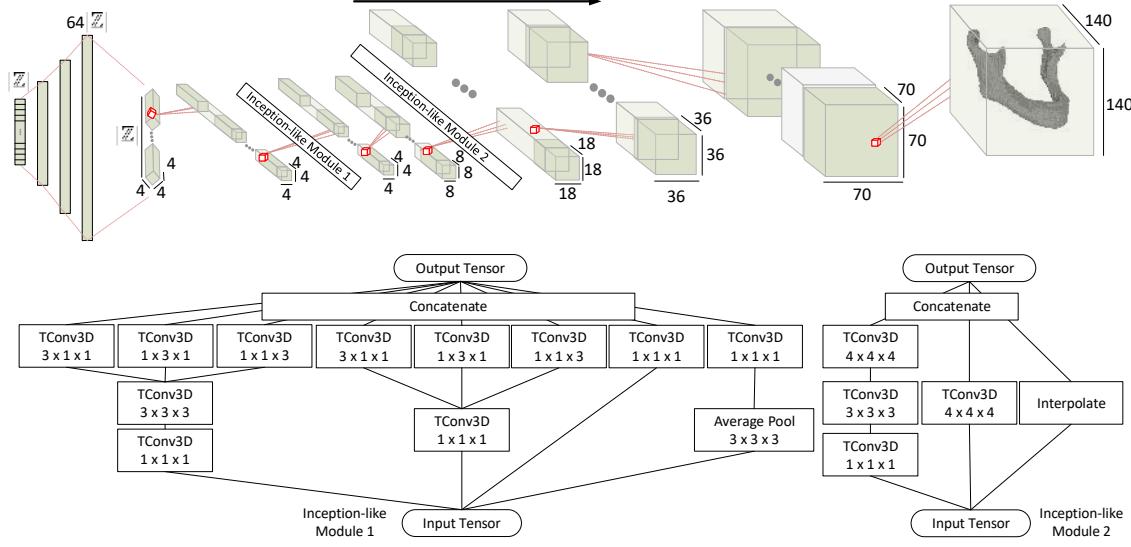


Figure 1: Architecture of the deep model for 3D shape generation from landmarks.

uously Differentiable Exponential Linear Units (CELU) with  $\alpha = 1$ , except for the last layer which is followed by a sigmoid function. Padding is applied when necessary to achieve the target voxel-grid cube of size  $140^3$ . The last TConv3D layer only has one kernel and its input is normalized using 3D batch normalization (BNorm3D).

Inspired by the success of the inception networks in designing deeper architectures for 2D image recognition tasks (Szegedy et al., 2016), 3D deconvolutional inception modules were designed and integrated into the AnatomyGen architecture. The inception-like blocks are depicted in Figure 1.

## 2.2. Loss Function

The generative model,  $g$ , is trained towards minimizing an approximated version of the Dice loss between the generated and expected voxel values, defined as follows

$$L_{dice}(\mathbf{V}, \hat{\mathbf{V}}) = \frac{2\langle \mathbf{V}, \hat{\mathbf{V}} \rangle_F}{\sum_{ijk} [\mathbf{V} + \hat{\mathbf{V}}]_{ijk}}, \quad (1)$$

where  $\langle \cdot \rangle_F$  is the Frobenius inner product of the two tensors.

## 2.3. Dense Representation

In general, the input space,  $\mathbb{Z}$ , can be any dense representation. In computer graphics, and in the context of 3D shape generation, a variety of choices have been investigated, including single or multiple 2D projections of the shape, shape silhouette, or vectors of classes or styles (Soltani et al., 2017; Firman et al., 2016).

In the context of 3D biomedical modeling, generating realistic shapes within the physiological variations of the population is valuable, yet, it is not necessarily coupled with clinical applications.

Description	Left No.	Right No.
Anterior point of the condylar process	1	16
Medial point of the condylar process	2	19
Posterior point of the condylar process	3	18
Lateral point of the condylar process	4	17
Deepest point of the mandibular notch	5	15
Peak of the coronoid process	6	14
Anterior end of the coronoid process	7	13
Mandibular foramen	8	12
Retromolar Fossa	9	11
Posterior concavity of the ramus	20	26
Mental tubercle	22	24
Angle of mandible	25	21
Anterior end of ramus	27	28
Genial tubercle on sagittal plane	10	
Mental protuberance on sagittal plane	23	

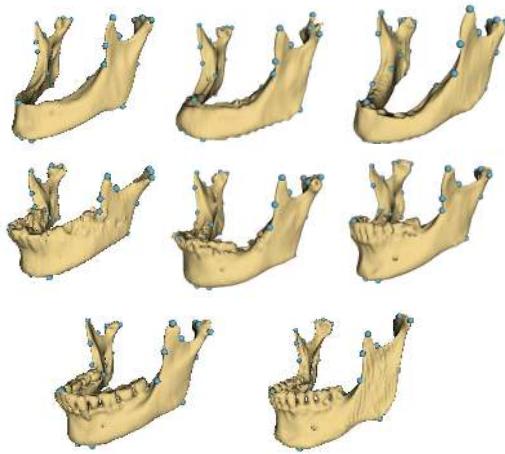


Figure 2: Some sample mandibles demonstrating the degree of variability across the data. The landmarks which constitute the input shape representation, along with their descriptions, are also visualized.

Generally speaking, there is a notion of clinical value and relevancy associated with certain regions of the anatomy. For instance, in biomechanical modeling, some regions are considered landmarks of substantial importance such as muscle and tendon insertion sites and collision surfaces. These areas play vital roles in force transmission and stress analysis; consequently, they are crucial in the validity and exactness of the resultant biomechanical model.

Based on the above intuition, and with guidance from an otolaryngology surgeon, 28 landmarks were selected on the surfaces of all samples. These landmarks were chosen based on two criteria: clinical relevancy, and reproducibility. Most of the landmarks corresponded to anatomical structures and muscle attachment sites (Figure 2). The validity of the landmarks was later checked by another dentist.

The 3D coordinates of the selected landmarks were used to form the input dense space for the generative model. As a result, the size of the input dense vector was  $|\mathbb{Z}| = 3n$ , where  $n$  is the total number of landmarks selected for each sample, in this case 28.

### 3. Data and Training

#### 3.1. Data

Our dataset includes a total of 103 mandibular meshes segmented from CT images. We collected the data from four publicly available data sources. The first subset of data comprises of 48 samples collected from the Vancouver General Hospital, Canada. Among the samples, 24 of them have teeth, while 24 mandibles are missing all or most of their dentition. The age range of the individuals was from 42 to 87. The second subset is published by Wallner et al. in figshare and consists of 10 mandibular samples (6 males, 4 females), all of whom lacked dentition (Wallner and Egger, 2018). The rest of the data were collected from the MICCAI Auto-segmentation challenge 2015 and The

Cancer Imaging Archive (TCIA) ([Raudaschl et al., 2017a; Zuley et al., 2016](#)). These mandibles were segmented as not to include the teeth and with a flat alveolar process.

### 3.2. Preprocessing

All 3D meshes were roughly aligned using the group-wise student's-t mixture model rigid registration algorithm based on their point clouds ([Ravikumar et al., 2016](#)). We used 50 mixture components for this alignment. Each closed surface mesh was then placed in the center of a 3D cubic voxel grid of size  $140^3$ . Each voxel was determined to be either inside or outside the mesh using ray tests and assigned a value of one or zero. This resulted in a discretized voxel-based representation of the mesh with isotropic voxels of size 1 mm, imitating a segmentation mask of the mandible obtained from an isotropic CT scan. The size of the matrix was set based on the maximum facial width reported in the comprehensive dataset of the FaceBase project ([Brinkley et al., 2016](#)).

### 3.3. Training

Adam optimizer, with default momentum parameters and  $\ell_2$  regularization of  $1e - 5$ , was used for training. The learning rate was initialized as  $1e - 3$ , with an exponential decay rate of 0.99 applied after every epoch. Weights were uniformly initialized from a symmetric interval defined inversely proportional to the number of input channels and kernel size to keep the output of the layer bounded within reasonable limits ([He et al., 2015](#)).

Dataset was randomly partitioned into 80% training-validation set, and 20% test set. Following common practices of data augmentation in convolutional networks, each training sample was randomly mirrored, shifted, and rotated. The probability of mirroring was 50% and was limited to the sagittal plane. The rotation was limited around the vertical axis. Each sample was augmented independently and the perturbed copies were generated on-the-fly during training. We used fixed random seeds for all the experiments to mitigate the inter-experimental variance.

The proposed network architecture was implemented using the open source PyTorch library. The implementation of the learning model and the training framework is publicly shared here: <https://github.com/amir-abdi/LandmarksToShape>. To enable reproducibility of the experiments, the preprocessed voxel-based representation of the data and their associated landmarks' coordinates accompanies the code according to each dataset's respective license and data sharing agreements.

## 4. Evaluation and Results

As opposed to the segmentation problem where the generated masks are compared with ground-truth manual annotations, there is no single ground-truth for a given input dense representation. In other words, since the generated shape is not constrained to comply with the edges of an image, the problem is ill-posed. Therefore, there is no easy way to evaluate the performance of the shape generation process.

In this work, the input abstract representation was formed by the 3D coordinates of surface landmarks. Thus, the average distance of these landmarks to the reconstructed surface (landmark-to-surface distance; L2S) is the most relevant metric for evaluation. For a model parameterized by  $\theta$ , the L2S metric for a set of landmarks  $z$  is calculated as

$$L2S(z, \theta) = \frac{1}{n} \sum_i^n d(z_i, g_\theta(z)), \quad (2)$$

Table 1: The performance of the trained model is evaluated based on the distance of the landmarks to the generated shape (L2S). The generated shapes were also compared with the test shapes, from which the landmarks were selected, based on the Dice Coefficient (DSC), Hausdorff at 95th percentile (HD95), and Surface Mean Distance (SMD) metrics. The results are compared with the average mandible shape and the best performing mandible segmentation models in the MICCAI challenge ([Raudaschl et al., 2017b](#)).

	L2S (mm)	HD95 (mm)	SMD (mm)	DSC (%)
MICCAI (Range)	—	2.5 - 10.5	0.5 - 2.8	78 - 93
Average Shape	$4.47 \pm 3.71$	$7.09 \pm 1.83$	$2.01 \pm 0.74$	$56.43 \pm 0.67$
<b>AnatomyGen</b>	<b><math>1.42 \pm 1.05</math></b>	<b><math>3.79 \pm 0.96</math></b>	<b><math>1.19 \pm 0.29</math></b>	<b><math>74.35 \pm 7.45</math></b>

where  $d(p, s)$  is the Euclidean distance of 3D point  $p$  to the surface of shape  $s$ . The L2S error for the landmarks selected on the samples of the test set was measured at  $1.42 \pm 1.05$  mm. Except for the landmarks on the peak of the coronoid processes (Figure 2, landmarks 6 & 14) with an error of close to 5 mm, the L2S error for almost all landmarks were below 2 mm ranging from 0.65 mm to 2.21 mm, with an average of  $1.16 \pm 0.46$  mm.

The test sample from which the landmarks were selected is one of the many possible solutions for the shape generation problem. Therefore, the generated samples were also compared with their corresponding reference test sample based on the Dice Coefficient (DSC), Surface Mean Distance (SMD) and Hausdorff at the 95th percentile (HD95). Finally, to make sure that the AnatomyGen model is taking the landmark coordinates into account, all the above metrics were also calculated for the average mandible shape. The results of the mentioned analyses are reported in Table 1. Some mandibles generated from unseen input dense vectors are visualized in Figure 3 for qualitative evaluation.

We should highlight that segmentation was not among the main objectives of the proposed method; however, it can still be used for semi-automated segmentation with manual selection of landmarks. Therefore, results from the best performing models of the MICCAI 2015 challenge in mandible segmentation are included in Table 1 only to demonstrate that the shapes generated by the proposed method are within reasonable standards.

To test the generalizability of the model, we ran experiments where we smoothly morphed one 3D shape into another by linearly interpolating in the  $\mathbb{Z}$ -space. Here, a set of landmarks selected on the surface of a test sample were incrementally updated towards landmarks of another test sample. At each increment, AnatomyGen was used to generate the mandible corresponding to the landmarks' coordinates. In Figure 4, the two test samples (left and right) and results of their linear interpolation are visualized.

## 5. Discussion and Conclusion

In this work, a deep 3D convolutional approach is proposed to generate anatomical shapes from dense vectorized representations. In our experiments, the 3D coordinates of a selected set of landmarks on the surface of the shape were used as the abstract representation of the shape. Thanks to the randomized and regularized training framework, the deep model was not only able to gen-



Figure 3: Mandibles generated using the AnatomyGen model from unseen vectors.

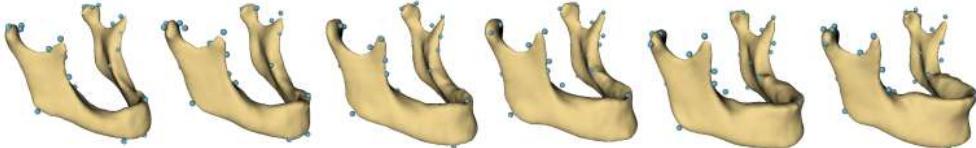


Figure 4: Smooth linear interpolation between two extreme cases of the test set.

eralize to unseen test samples and generate realistically looking samples but also demonstrated its potential in interpolating in-between shapes. The average distances between input landmarks and the generated surfaces (L2S) were close to 1 mm, which is similar to the resolution of the voxel grid.

The only exceptions to the high-performing L2S metric were the landmarks on the peak of the coronoid processes (Figure 2). Our hypothesis is that the coronoid processes were so thin that the errors in their reconstruction did not contribute much to the Dice loss. As a result, the model was not highly penalized for inaccurate reconstruction of this thin structure. Including the L2S metric in the loss function would have mitigated this error; however, it would have limited the generalizability of the proposed method as we had aimed for the input to be any abstract representation and not be limited to landmarks.

The dataset used in this study were formed by merging four data sources; thus, they were diversified with respect to their anatomical variations, the presence of teeth, and the amount of bone loss.

Moreover, the segmentations were carried out by various experts on images of different imaging systems which resulted in the occasional presence of holes and gaps in a subset of samples due to their imprecise segmentations.

No information beyond the input dense vector is provided to the AnatomyGen model. As a result, the proposed shape generation process is expected to be a one-to-many mapping between the encoded dense representation and the 3D voxel-based shape. Consequently, if the trained model generates a shape from a set of landmarks selected on the surface of a given shape, the two anatomies will not necessarily be identical. However, during model evaluation, the samples of the test set, from which the landmarks originated, were considered as one of the many acceptable potential reconstructions. With this assumption in mind, we compared the generated mandibles with their corresponding sample of the test set (Table 1).

A limitation of this study lies in its fully supervised training without encoding the posterior distribution; thus, there is no principled way for this model to generate random samples. However, AnatomyGen demonstrated acceptable interpolation potential in incrementally morphing a given shape to another (Figure 4). As a result, a possible solution for random shape synthesis is to model the statistical coordinates of the landmarks as a multivariate distribution and sample from to generate random mandibles.

A potential application of the AnatomyGen model is in semi-automated segmentation. In this approach, the expert manually selects a small number of predefined landmarks on the 3D medical image and the deep model generates the corresponding segmentation map. This approach is modality agnostic as no image features are included in the process.

The approach introduced here differs from statistical shape models as it does not morph a model across modes of variation, but generates an anatomy from an abstract representation. It is also different from segmentation methods as the generative model is blind to the image features. Basically, the proposed method is a step towards generating biomedical models from abstract representations.

Our next step is to investigate shape generation and reconstruction from other representative forms, such as partial shapes, where the missing parts of the anatomy are completed using a deep generative model. This approach is helpful in surgical planning where the normal pre-morbid mandibular form is unknown.

## Acknowledgments

We would like to thank Dr. Eitan Prisman from the Vancouver General Hospital for his support throughout this research. This research was undertaken, in part, thanks to the funding from the Vanier Scholar award of the Natural Sciences and Engineering Research Council of Canada (NSERC) to the first author, Amir H. Abdi.

## References

- Michael Skipper Andersen, Mark De Zee, Michael Damsgaard, Daniel Nolte, and John Rasmussen. Introduction to Force-Dependent Kinematics: Theory and Application to Mandible Modeling. *Journal of Biomechanical Engineering*, 2017.
- Christoph Baur, Shadi Albarqouni, and Nassir Navab. Generating highly realistic images of skin lesions with GANs. In *Lecture Notes in Computer Science*, pages 260–267. Springer International Publishing, 2018.

James F Brinkley, Shannon Fisher, Matthew P Harris, Greg Holmes, Joan E Hooper, Ethylin Wang Jabs, Kenneth L Jones, Carl Kesselman, Ophir D Klein, Richard L Maas, Mary L Marazita, Licia Selleri, Richard A Spritz, Harm van Bakel, Axel Visel, Trevor J Williams, Joanna Wysocka, FaceBase FaceBase Consortium, and Yang Chai. The FaceBase Consortium: a comprehensive resource for craniofacial researchers. *Development*, 143(14):2677–88, 2016. ISSN 1477-9129. URL <https://www.facebase.org>.

Andrew Brock, Theodore Lim, J. M. Ritchie, and Nick Weston. Generative and discriminative voxel modeling with convolutional neural networks. In *Neural Information Processing Conference*, 2016.

Christopher B. Choy, Danfei Xu, Jun Young Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *Lecture Notes in Computer Science*, volume 9912 LNCS, pages 628–644, 2016.

Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J. Brostow. Structured Prediction of Unobserved Voxels from a Single Depth Image. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5431–5440, 2016.

Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *Lecture Notes in Computer Science*, volume 9910 LNCS, pages 484–499, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015.

Jennifer L. Hicks, Thomas K. Uchida, Ajay Seth, Apoorva Rajagopal, and Scott L. Delp. Is My Model Good Enough? Best Practices for Verification and Validation of Musculoskeletal Models and Simulations of Movement. *Journal of Biomechanical Engineering*, 137(2):020905, 2015.

Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. Gans for medical image analysis. *Computing Research Repository (CoRR)*, abs/1809.06222, 2018.

Jerry Liu, Fisher Yu, and Thomas Funkhouser. Interactive 3D Modeling with a Generative Adversarial Network. In *International Conference on 3D Vision (3DV 2017)*, pages 126–134, 2017.

Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A. Cook, Antonio de Marvao, Timothy Dawes, Declan P. Regan, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation. *IEEE Transactions on Medical Imaging*, 37(2):384–395, 2018.

Yongsheng Pan, Mingxia Liu, Chunfeng Lian, Tao Zhou, Yong Xia, and Dinggang Shen. Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer’s disease diagnosis. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2018)*, volume 11072 LNCS, pages 455–463. Springer, Cham, 2018.

Patrik F. Raudaschl, Paolo Zaffino, Gregory C. Sharp, Maria Francesca Spadea, Antong Chen, Benoit M. Dawant, Thomas Albrecht, Tobias Gass, Christoph Langguth, Marcel Lüthi, Florian Jung, Oliver Knapp, Stefan Wesarg, Richard Mannion-Haworth, Mike Bowes, Annaliese Ashman, Gwenael Guillard, Alan Brett, Graham Vincent, Mauricio Orbes-Arteaga, David Cárdenas-Peña, German Castellanos-Dominguez, Nava Aghdasi, Yangming Li, Angelique Berens, Kris Moe, Blake Hannaford, Rainer Schubert, and Karl D. Fritscher. Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015. *Medical Physics*, 44(5): 2020–2036, 2017a.

Patrik F. Raudaschl, Paolo Zaffino, Gregory C. Sharp, Maria Francesca Spadea, Antong Chen, Benoit M. Dawant, Thomas Albrecht, Tobias Gass, Christoph Langguth, Marcel Lüthi, Florian Jung, Oliver Knapp, Stefan Wesarg, Richard Mannion-Haworth, Mike Bowes, Annaliese Ashman, Gwenael Guillard, Alan Brett, Graham Vincent, Mauricio Orbes-Arteaga, David Cárdenas-Peña, German Castellanos-Dominguez, Nava Aghdasi, Yangming Li, Angelique Berens, Kris Moe, Blake Hannaford, Rainer Schubert, and Karl D. Fritscher. Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015. *Medical Physics*, 44(5): 2020–2036, 2017b. ISSN 00942405.

Nishant Ravikumar, Ali Gooya, Serkan Çimen, Alejandro F. Frangi, and Zeike A. Taylor. A multi-resolution T-mixture model approach to robust group-wise alignment of shapes. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2016)*, volume 9902 LNCS, pages 142–149. Springer, Cham, 2016.

Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. Variational approaches for auto-encoding generative adversarial networks. *Computing Research Repository (CoRR)*, abs/1706.04987, 2017.

Edward J. Smith and David Meger. Improved adversarial systems for 3d object generation and reconstruction. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 87–96. PMLR, 2017.

Amir Arsalan Soltani, Haibin Huang, Jiajun Wu, Tejas D. Kulkarni, and Joshua B. Tenenbaum. Synthesizing 3D Shapes via Modeling Multi-view Depth Maps and Silhouettes with Deep Generative Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2511–2519. IEEE, 2017.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017.

- Jürgen Wallner and Jan Egger. Mandibular CT dataset collection. figshare, 2018.
- Jelmer M Wolterink, Peter R Seevinck, and Anna M Dinkla. MR-to-CT Synthesis using Cycle-Consistent Generative Adversarial Networks. In *Med-NIPS*, 2017.
- Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T. Freeman, and Joshua B. Tenenbaum. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. *Graphical Models and Image Processing (CVGIP)*, 53(2):157–185, 2016.
- Miaomiao Zhang and Polina Golland. Statistical shape analysis: From landmarks to diffeomorphisms. *Medical Image Analysis*, 33:155–158, 2016.
- Margarita L. Zuley, Rose Jarosz, Shanah Kirk, Yueh Lee, Rivka Colen, Kimberly Garcia, Dominique Delbeke, Michelle Pham, Paul Nagy, Gorkem Sevinc, Marla Goldsmith, Subair Khan, Jose M. Net, Fabiano R. Lucchesi, and Natalia D. Aredes. Radiology data from the cancer genome atlas head-neck squamous cell carcinoma collection. The Cancer Imaging Archive, 2016.

# Exploring local rotation invariance in 3D CNNs with steerable filters

**Vincent Andrearczyk<sup>1</sup>**

**Julien Fageot<sup>2</sup>**

**Valentin Oreiller<sup>1,3</sup>**

**Xavier Montet<sup>4</sup>**

**Adrien Depeursinge<sup>1,3</sup>**

<sup>1</sup> University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland

<sup>2</sup> Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

<sup>3</sup> Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland

<sup>4</sup> Hopitaux Universitaires de Genève (HUG), Geneva, Switzerland

## Abstract

Locally Rotation Invariant (LRI) image analysis was shown to be fundamental in many applications and in particular in medical imaging where local structures of tissues occur at arbitrary rotations. LRI constituted the cornerstone of several breakthroughs in texture analysis, including Local Binary Patterns (LBP), Maximum Response 8 (MR8) and steerable filterbanks. Whereas globally rotation invariant Convolutional Neural Networks (CNN) were recently proposed, LRI was very little investigated in the context of deep learning. We use trainable 3D steerable filters in CNNs in order to obtain LRI with directional sensitivity, i.e. non-isotropic filters. Pooling across orientation channels after the first convolution layer releases the constraint on finite rotation groups as assumed in several recent works. Steerable filters are used to achieve a fine and efficient sampling of 3D rotations. We only convolve the input volume with a set of Spherical Harmonics (SHs) modulated by trainable radial supports and directly steer the responses, resulting in a drastic reduction of trainable parameters and of convolution operations, as well as avoiding approximations due to interpolation of rotated kernels. The proposed method is evaluated and compared to standard CNNs on 3D texture datasets including synthetic volumes with rotated patterns and pulmonary nodule classification in CT. The results show the importance of LRI in CNNs and the need for a fine rotation sampling.

**Keywords:** Local rotation invariance, convolutional neural network, steerable filters, 3D texture

## 1. Introduction

Convolutional Neural Networks (CNNs) have been used in various studies to analyze textures. Orderless pooling of feature maps is used to discard the overall shape and layout information and, thus, describe repetitive and diffuse texture patterns (Andrearczyk and Whelan, 2016; Cimpoi et al., 2016; Zhang et al., 2016). By construction, CNN architectures provide translation equivariance, which is particularly adapted to image analysis. This paper focuses on adding local rotation invariance in the CNN architecture, which is known to be crucial for biomedical applications (Depeursinge and Fageot, 2018).

Globally rotation equivariant/invariant CNNs have recently been extensively studied using group theory in order to propagate rotation equivariance throughout the network. The 2D Group equivariant CNNs (G-CNN) introduced in (Cohen and Welling, 2016) uses rotated convolutional filters

with right angle rotations of the  $p4$  symmetry group. Invariance is obtained by pooling across orientation channels after the last convolution layer. The G-CNN was recently extended to 3D images in (Winkels and Cohen, 2018) showing a performance increase in the analysis of pulmonary nodule detection. 3D G-CNNs were shown to improve classification of 3D textures (Andrarczyk and Depeursinge, 2018), yet the results motivated the use of a finer rotation sampling than right angle rotations from the Octahedral  $O$  group to capture realistic arbitrary 3D orientations of directional patterns. It is important to remark that G-CNNs are adapted to equivariance with respect to *finite* subgroups of the rotation group. In 2D, an arbitrary sampling of rotations can be used (Bekkers et al., 2018; Zhou et al., 2017) in a group equivariant approach, while the number of 3D finite rotation groups is restrained. The 2D harmonic network (Worrall et al., 2016) and 2D steerable CNN (Weiler et al., 2017) present similarities with the method proposed in this paper although in the 2D domain and not particularly designed for texture analysis. Finally, the 3D steerable CNNs (Weiler et al., 2018) are very general architectures that implement the global equivariance to rotation on the network, and the convolutional layer considered in this paper is covered by their characterization. As detailed below, we differ from their works by making an angular max pooling after the first convolution layer, which exploits the steerability of the filters, and more importantly, focuses on local invariances.

In the above approaches, global rotation equivariance is maintained all along the layers (see Fig. 1, left), and invariance is obtained by using orientation pooling at the end of the network after spatial average pooling. Global rotation invariance is fundamental in various applications. However, some images are composed of well-defined structures with arbitrary orientations. For instance, 3D textures observed in Computed Tomography (CT) and in Magnetic Resonance Imaging (MRI) exhibit diverse tissue alterations, including necrosis, angiogenesis, fibrosis, or cell proliferation (Gatenby et al., 2013). These alterations induce imaging signatures such as blobs, intersecting surfaces and curves. These local low-level patterns are characterized by discriminative directional properties and have arbitrary 3D orientations, which requires combining directional sensitivity with LRI. How-

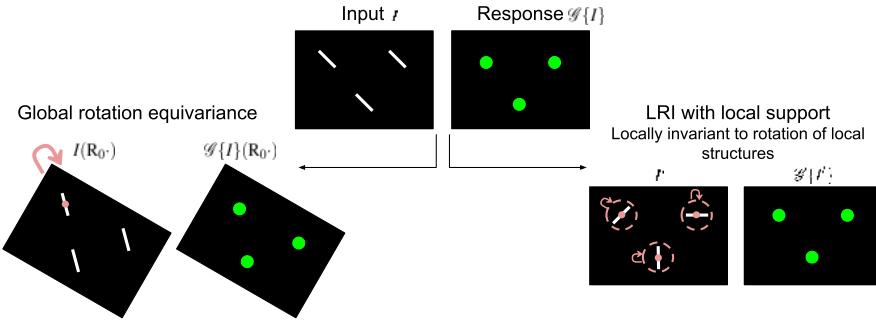


Figure 1: Illustration of global rotation equivariance and LRI in 2D. Rotating local structures (i.e. three white segments) in the input  $I$  results in the input  $I'$  on the right. The green dots illustrate the equivariant/invariant responses. Local and global rotations are shown in red and the local support  $G$  of the operator  $\mathcal{G}$  (see Section 2.1) is represented as a dashed red line. It is worth noting that our CNN architecture will both present a global equivariance and a local invariance to rotations. Best viewed in color.

ever, rotation invariance is often antagonist with the will of being sensitive to directional features. The latter is required to avoid mixing blobs, edges and ridges. For instance, a spatial convolutional operator is equivariant to rotations if and only if the filter is isotropic, therefore insensitive to the directional features of the input signal. It follows that operators combining LRI and directional sensitivity require using more complex designs such as MR8 (Varma and Zisserman, 2005), local binary patterns (Ojala et al., 2002), 3D Riesz wavelets (Dicente Cid et al., 2017) and Spherical Harmonic (SH) invariants (Depeursinge et al., 2018) widely used in hand-crafted texture analysis (Depeursinge and Fageot, 2018).

In this paper, we exploit the steerability of SHs to obtain a CNN architecture which is both globally equivariant and locally invariant to rotations (see Fig. 1 for a 2D illustration). This is achieved with a fine rotation sampling and controlled operator support. The local support for the rotation invariance is set by the kernel size of the first layer. LRI is then obtained by pooling across orientations after this first layer. The implementation will be made publicly available.

## 2. Methods

We first introduce the framework in the continuous domain, hence voxel images, filters, and response maps are functions defined over the continuum  $\mathbb{R}^3$ . The discretization is then presented in Section 2.4. Spherical coordinates are defined as  $(\rho, \theta, \phi)$  with radius  $\rho \geq 0$ , elevation angle  $\theta \in [0, \pi]$ , and horizontal plane angle  $\phi \in [0, 2\pi]$ . The set of 3D rotations is denoted by  $SO(3)$ . A 3D rotation transformation  $R$  can be decomposed as three elementary rotations around the  $z$ ,  $y'$  and  $z''$  axes as  $R = R_\alpha R_\beta R_\gamma$ , with the (intrinsic) Euler angles  $\alpha \in [0, 2\pi]$ ,  $\beta \in [0, \pi]$ , and  $\gamma \in [0, 2\pi]$  respectively. We will use interchangeably  $R$  as a rotation transformation acting on  $\mathbb{R}^3$  and on the two-dimensional sphere  $\mathbb{S}^2$ . Finally, the function  $x \mapsto f(Rx)$  is denoted by  $f(R\cdot)$ .

### 2.1. Equivariant Local Texture Operators

We introduce the class of texture operators of interest that will be used in the first layer of our neural network. We consider a filter  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ , whose support  $G$  is assumed to be finite. For an image  $I$  and a position  $x \in \mathbb{R}^3$ , we define the operator

$$\mathcal{G}\{I\}(x) = \max_{R \in SO(3)} |(I * f(R\cdot))(x)|. \quad (1)$$

The operator combines a convolutional operator together with a max-pooling operation over the rotations  $R$ , and is an example of texture operator as presented in (Depeursinge and Fageot, 2018). Note that  $(I * f(R\cdot))(x)$  can be identified as a lifting group-convolution (Bekkers et al., 2018), followed by sub-group pooling (maximum over rotations). Then,  $\mathcal{G}$  has the following properties:

- It is *globally equivariant to translations and rotations*, in the sense that, for any position  $x_0 \in \mathbb{R}^3$  and rotation  $R_0 \in SO(3)$ ,

$$\mathcal{G}\{I(\cdot - x_0)\} = \mathcal{G}\{I\}(\cdot - x_0) \text{ and } \mathcal{G}\{I(R_0 \cdot)\} = \mathcal{G}\{I\}(R_0 \cdot). \quad (2)$$

The proof is provided in Appendix A. In particular, if  $R_{x_0}$  is a rotation around  $x_0 \in \mathbb{R}^3$ , we have that  $\mathcal{G}\{I(R_{x_0} \cdot)\} = \mathcal{G}\{I\}(R_{x_0} \cdot)$ , as illustrated on the left part of Fig. 1.

- It is *local* in the sense that the filter  $f$  has a finite support  $G = \{x \in \mathbb{R}^3, \|x\| \leq \rho_0\}$ . As a consequence,  $\mathcal{G}\{I\}(x)$  only depends on the values  $I(y)$  for  $\|y - x\| \leq \rho_0$ .

The global equivariance to translations and rotations together with the locality create an invariance to local rotations (i.e. LRI) in the following sense: the rotation of an object or localized structure of interest in the image  $I$  around a position  $\mathbf{x}$  does not affect the value of  $\mathcal{G}\{I\}(\mathbf{x})$ , as illustrated on the right part of Fig. 1.

## 2.2. Steerable Filters and Spherical Harmonics

Computing the texture operator (1) requires to maximize over any 3D rotation  $R$  for every position  $\mathbf{x}$  of the image  $I$ , which can be computationally discouraging. To overcome this issue, we propose to use steerable filters, which have the advantage to allow for fast and efficient max pooling rotations (Chenouard and Unser, 2012; Fageot et al., 2018). A filter is steerable if any of its rotated version can be written as a linear combination of finitely many basis filters (Freeman and Adelson, 1991; Unser and Chenouard, 2013).

We consider filters  $f$  that are polar-separable, in the sense that they can be written as  $f(\rho, \theta, \phi) = h(\rho)g(\theta, \phi)$  with  $h : \mathbb{R}^+ \rightarrow \mathbb{R}$  and  $g : \mathbb{S}^2 \rightarrow \mathbb{R}$ . One can expand such steerable polar-separable filters in terms of the family of SHs  $(Y_{n,m})_{n \geq 0, m \in \{-n, \dots, n\}}$ , where  $n$  is called the degree and  $m$  the order, and which form an orthonormal basis for square-integrable functions  $g(\theta, \phi)$  on  $\mathbb{S}^2$ . We consider finitely many degrees,  $N \geq 0$  being the maximal one. In particular, the number of elements of a SH family of maximum degree  $N$  is  $\sum_{n=0}^N (2n+1) = (N+1)^2$ . The definition of SHs can be found in Appendix B.

The general form of a polar-separable steerable filter with maximal degree  $N \geq 0$  is

$$f(\rho, \theta, \phi) = h(\rho)g(\theta, \phi) = h(\rho) \sum_{n=0}^N \sum_{m=-n}^n C_n[m] Y_{n,m}(\theta, \phi), \quad (3)$$

where  $h(\rho)$  is the radial profile of  $f$  and the coefficients  $C_n[m]$  determine the angular profile  $g(\theta, \phi)$ . The condition of  $f$  being real is translated into the conditions that  $h$  itself is real and that the SH coefficients satisfy  $C_n[-m] = (-1)^m \overline{C_n[m]}$  (see Appendix C).

For any  $R \in SO(3)$ , the rotated version  $Y_{n,m}(R \cdot)$  of a SH can be expressed as

$$Y_{n,m}(R \cdot) = \sum_{m'=-n}^n D_{R,n}[m, m'] Y_{n,m'}. \quad (4)$$

where the  $D_{R,n} \in \mathbb{C}^{(2n+1) \times (2n+1)}$  are the Wigner matrices (Varshalovich et al., 1988). Then, the steerable filter  $f$  can then be rotated efficiently with any  $R \in SO(3)$  to obtain a set of steered coefficients  $C_{R,n} = D_{R,n} C_n$  of  $f(R \cdot)$ , with  $C_n = (C_n[m])_{m \in \{-n, \dots, n\}}$ . Then, the rotated filter  $f(R \cdot)$  is given by

$$f(R \cdot)(\rho, \theta, \phi) = h(\rho) \sum_{n=0}^N \sum_{m=-n}^n \sum_{m'=-n}^n D_{R,n}[m, m'] C_n[m'] Y_{n,m}(\theta, \phi). \quad (5)$$

From (5), we see that any rotated version of  $f$  can be computed from the coefficients  $(C_n[m])_{0 \leq n \leq N, -n \leq m \leq n}$ .

## 2.3. 3D Steerable Convolution and Max Pooling

Exploiting (5), the convolutional operator  $I * f(R \cdot)$  in (1) is then computed as

$$I * f(R \cdot) = \sum_{n=0}^N \sum_{m=-n}^n \left( \sum_{m'=-n}^n D_{R,n}[m, m'] C_n[m'] \right) (I * h Y_{n,m}). \quad (6)$$

Therefore, one accesses the convolution with any rotated version of  $f$  by computing  $\sum_n (2n+1) = (N+1)^2$  convolutions ( $I * hY_{n,m}$ ), which we shall exploit for computing the response map  $\mathcal{G}\{I\}$  of the texture operator (1). It is worth noting that the case  $N=0$  corresponds to filters  $f$  that are isotropic, i.e.  $f(\mathbf{R}\cdot) = f$  for any  $\mathbf{R} \in SO(3)$  (Depeursinge et al., 2018). As low degrees (e.g.  $N=1, 2$ ) are sufficient to construct small filters (see Section 2.4), the gain becomes substantial over a G-CNN approach for a fine sampling of orientations with a drastic reduction of the number of convolutions.

In practice, one has a set of steerable filters  $f_i$  of the form (3) with radial profiles  $h_i$  and coefficients  $C_{i,n}[m]$ . The number of trainable parameters is reduced to the coefficients  $C_{i,n}[m]$ , the radial profiles  $h_i$  and the biases (one scalar parameter per output channel  $i$ ).

## 2.4. Discretization

The radial profiles  $h_i$ , and hence the filters  $f_i$ , have a compact spherical support  $G = \{\mathbf{x} \in \mathbb{R}^3, \|\mathbf{x}\| \leq \rho_0\}$ , where  $\rho_0 > 0$  is fixed. For any  $i$ , we consider the voxelized version of the radial profile  $h_i(\rho)$ . The size of the support of the voxelized version is linked to the radius  $\rho_0$  of the filter in the continuous domain and the level of voxelization. Due to the isotropic constraint, for a support of  $c^3$  voxels, the number of trainable parameters for each  $h_i$  is then  $\left\lceil \frac{(c-1)}{2} \times \sqrt{3} \right\rceil + 1$ . The values of the filter  $f_i(\rho, \theta, \phi)$  over the continuum is deduced from the discretization using linear interpolation (Fig. 2). Note that the corner effect (radial values for  $\rho > \frac{(c-1)}{2}$ ) has a limited impact on the approximated rotation invariance and the freedom is maintained for these parameters to be set to zero during training.

The maximal frequency  $N$  cannot be taken arbitrarily large once the radial profiles are voxelized (Weiler et al., 2017). Indeed, the discretized filters  $f_i$  are defined over  $c^3$  voxels, which imposes the upper bound  $N \leq \rho_0 c / 2$ , where  $\rho_0$  is the radius of the spherical support of  $h_i$ . This can be interpreted as the angular Nyquist frequency.

To sample the rotations, we uniformly sample points on the sphere using a triangulation method that iteratively splits octahedron faces to obtain the  $(\alpha, \beta)$  Euler angles around  $z$  and  $y'$  respectively. We then sample the last angle  $\gamma$  around  $z''$  uniformly between 0 and  $2\pi$ . The Octahedral group, for instance, is obtained by sampling 6 points on the sphere (i.e. six  $(\alpha, \beta)$  pairs) and four values of  $\gamma$  to obtain 24 right angle rotations. In this paper, we evaluate the following sets of rotations: single rotation, Klein's four rotations, octahedral 24 rotations and 72 rotations (18 points on the sphere and 4 values of  $\gamma$ ). In the sequel, we denote by  $M$  the number of tested rotations.

## 2.5. Datasets

We evaluate the proposed method on two experiments described in the following. In the first experiment, we built a dataset containing two classes of 500 synthetic volumes each. The volumes of

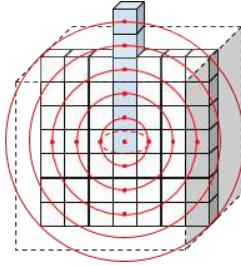


Figure 2: Illustration of a 2D slice of the isotropic radial profile  $h_i$ . The blue voxels represent the trainable parameters. The rest of the cube is linearly interpolated.

size  $32 \times 32 \times 32$  are generated by placing two  $7 \times 7 \times 7$  patterns, namely a binary segment and a 2D cross with the same norm, at random 3D orientations and random locations with overlap. The number of patterns per volume is randomly set to  $\lfloor d(\frac{s_v}{s_p})^3 \rfloor$ , where  $s_v$  and  $s_p$  are the sizes of the volume and of the pattern respectively and the density  $d$  is in the range [0.2, 0.4]. The two classes vary by the proportion of the patterns, i.e. 10% segments with 90% crosses for the first class and vice versa for the second class. 800 volumes are used for training and the remaining 200 for testing. Despite the simplicity of this dataset, some variability is introduced by the overlapping patterns and the linear interpolation of the 3D rotations, making it challenging and more realistic.

The second dataset is a subsample of the American National Lung Screening Trial (NLST) that was annotated by radiologists at the University Hospitals of Geneva (HUG) ([Martin et al., submitted](#)). The dataset comprises 485 pulmonary nodules from distinct patients in CT, among which 244 were labeled benign and 241 malignant. We pad or crop the input volumes (originally ranging from  $16 \times 16 \times 16$  to  $128 \times 128 \times 128$ ) to the size  $64 \times 64 \times 64$ . We use the balanced training and test splits with 392 and 93 volumes respectively. Examples of 2D slices of the lung nodules are illustrated in Fig. 3. The Hounsfield units are clipped in the range  $[-1000, 400]$ , then normalized with zero mean and unit variance (using the training mean and variance).

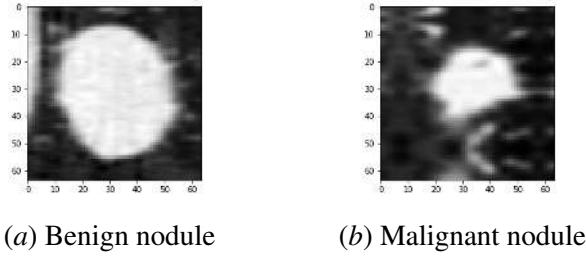


Figure 3: 2D slices from 3D volumes of benign and malignant pulmonary nodules.

## 2.6. Network Architecture

The first layer of the networks is the LRI layer (1). Global average spatial pooling is then used similarly to ([Andrarczyk and Whelan, 2016](#)). This pooling aggregates the locally invariant texture responses into a single scalar per feature map and is followed by fully connected layers. For the nodule experiment, we average the responses inside the nodule masks instead of across the entire feature maps. For the synthetic experiment, we connect directly the final softmax fully connected layer with a cross-entropy loss. For the second, more complex experiment, we use an intermediate fully connected layer with 128 neurons before the same final layer. Standard ReLU activations are employed. The networks are trained using Adam optimizer with  $\beta_1 = 0.99$  and  $\beta_2 = 0.9999$  and a batch size of 8. Other task-specific parameters are: for the synthetic experiment (kernel size  $7 \times 7 \times 7$ , stride 1, 2 filters and 50,000 iterations), for the nodule experiment (kernel size  $9 \times 9 \times 9$ , stride 2, 4 filters and 10,000 iterations).

We refer to the developed architecture as LRI-CNN and compare it to a network with the same architecture but with a standard 3D convolution layer, referred to as Z3-CNN.

## 2.7. Weights Initialization

The SHs are normalized to  $\|Y_{n,m}\|_2 = 1$ . The coefficients are then randomly initialized by a normal distribution with  $Var[C_{i,n}[m]] = \frac{2}{n_{in}(N+1)^2}$ , where  $n_{in}$  is the number of input channels (generally 1), the radial profiles are initialized to  $Var[h_i(\rho)] = 1$  and the biases to zero. This initialization is inspired by (He et al., 2015; Weiler et al., 2017) in order to avoid vanishing and exploding activations and gradients.

## 3. Experimental Results

The results for the synthetic experiment (3D textures of synthetic rotated patterns) are summarized in Table 1. Fig. 4 shows a comparison of standard 3D kernels (Z3-CNN) and SH parametric representations (LRI-CNN).

Table 1: Average accuracy (%) on the synthetic 3D local rotation dataset with  $N = 2$ .

model	# orient. (M)	# filters	# param.	accuracy $\pm\sigma$
Z3-CNN	-	2	694	81.7 $\pm$ 4.4
Z3-CNN	-	144	49,826	96.0 $\pm$ 0.3
LRI-CNN	1	2	40	74.6 $\pm$ 3.2
LRI-CNN	4	2	40	85.4 $\pm$ 4.7
LRI-CNN	24	2	40	88.2 $\pm$ 2.9
LRI-CNN	72	2	40	90.0 $\pm$ 1.3

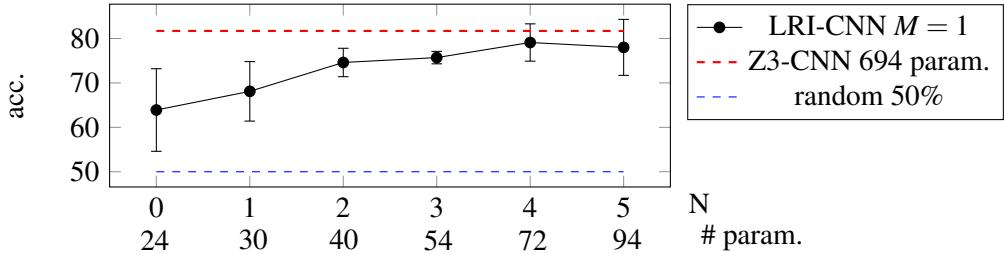


Figure 4: Comparison of standard 3D kernels (Z3-CNN) and SH parametric representation (LRI-CNN) with varying maximum degree  $N$  with a single orientation  $M = 1$  (i.e. not using the steering capacity) on the synthetic 3D local rotation dataset. When  $N \geq 2$ , the performance of the SH parametric representation is very close to the Z3-CNN while using  $15\times$  fewer parameters.

Results for the nodule classification experiment on the pulmonary nodules classification (NLST) are summarized in Table 2. The results are averaged over 10 repetitions.

Table 2: Average accuracy (%) on the pulmonary nodule classification with  $N = 2$ .

model	# orient. (M)	# filters	# param.	accuracy $\pm \sigma$
Z3-CNN	-	4	3,818	80.0 $\pm$ 1.7
Z3-CNN	-	96	82,754	81.3 $\pm$ 2.2
LRI-CNN	1	4	970	76.3 $\pm$ 3.8
LRI-CNN	4	4	970	79.0 $\pm$ 3.0
LRI-CNN	24	4	970	<b>81.9<math>\pm</math>3.3</b>
LRI-CNN	72	4	970	80.7 $\pm$ 7.4

#### 4. Discussions and Conclusion

The results on the synthetic dataset (Table 1) show that increasing the number of orientation channels significantly improves the performance (74.6% with a single orientation vs 90.0% with 72 orientations) and outperforms a standard Z3-CNN with the same number of filters (81.7%). Despite the increased number of orientation channels, the number of trainable parameters remains extremely low (40 parameters). Note that using data augmentation with random rotations of the training samples would not help the Z3-CNN as its architecture is simply inappropriate for LRI and patterns are already present at many random orientations in the training set. Adding more filters to the Z3-CNN ( $2 \times 72 = 144$  filters) allows to learn filters at different orientations and achieves 95.9% accuracy, at the heavy cost of parameters and convolution operations. As shown in Fig. 4 with a single orientation channel, i.e. without using the steering capacity, the degree  $N = 0$  of the SH cannot differentiate well patterns (63.9% accuracy) as it is isotropic. The performance then increases with  $N$  and nearly reaches the standard Z3-CNN accuracy for  $N \geq 2$  with a significantly lower number of parameters, underlining the compression power of the parametric SH representation.

Note that LRI can be obtained with a G-CNN implementation (Cohen and Welling, 2016) by pooling across orientation channels after the first layer, yet it is limited to M=24 and requires to convolve the input with every rotated filter.

The results on the pulmonary nodule classification experiment (see Table 2) confirm the importance of LRI and of the proposed approach in a real medical imaging application. Despite the lack of directional texture patterns in the data which may reduce the performance gain over the baseline, an increase in accuracy is obtained with the LRI-CNN as well as a reduction of trainable parameters by a factor of four.

In conclusion, we developed a 3D LRI convolutional network using steerable filters. The main benefits are the low number of trainable parameters, the limited number of convolutions as we only convolve with the limited set of SHs and steer the responses for an arbitrary number of rotations, and the exactness of the steerability, avoiding approximation for kernel rotations. The results on synthetic 3D textures and 3D pulmonary nodule classification confirmed the importance of LRI with directionally sensitive steerable filters and the compression power of the proposed approach. While the reduction of trainable parameters also comes with an increase of time and memory as compared with a standard 3D CNN, the proposed efficient and precise LRI design may be preferred over data augmentation or group equivariant networks in various scenarios. In future work, we will look into finding the maximum orientation responses and/or powerful invariant descriptors without recombining the responses for all orientations which is a current bottleneck for memory

consumption on the GPUs. We will also explore the benefit and cost of using non-polar-separable filters.

## Acknowledgments

This work was supported by the Swiss National Science Foundation (grant 205320\_179069).

## References

- M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, volume 55. Courier Corporation, 1964.
- V. Andrearczyk and A. Depeursinge. Rotational 3D texture classification using group equivariant CNNs. *arXiv preprint arXiv:1810.06889*, 2018.
- V. Andrearczyk and P.F. Whelan. Using filter banks in convolutional neural networks for texture classification. *Pattern Recognition Letters*, 84:63–69, 2016.
- E. J. Bekkers, M. W. Lafarge, M. Veta, K. AJ. Eppenhof, J. PW Pluim, and R. Duits. Roto-translation covariant convolutional networks for medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 440–448. Springer, 2018.
- N. Chenouard and M. Unser. 3D steerable wavelets in practice. *IEEE Transactions on Image Processing*, 21(11):4522–4533, 2012.
- M. Cimpoi, S. Maji, I Kokkinos, and A. Vedaldi. Deep filter banks for texture recognition, description, and segmentation. *International Journal of Computer Vision*, 118(1):65–94, 2016.
- T.S. Cohen and M. Welling. Group equivariant convolutional networks. *CoRR*, abs/1602.07576, 2016.
- A. Depeursinge and J. Fageot. Biomedical texture operators and aggregation functions: A methodological review and user’s guide. In *Biomedical Texture Analysis*, pages 55–94. Elsevier, 2018.
- A. Depeursinge, J. Fageot, V. Andrearczyk, J.P. Ward, and M. Unser. Rotation invariance and directional sensitivity: Spherical harmonics versus radiomics features. In *International Workshop on Machine Learning in Medical Imaging*, pages 107–115. Springer, 2018.
- Y. Dicente Cid, H. Müller, A. Platon, P.A. Poletti, and A. Depeursinge. 3-D solid texture classification using locally-oriented wavelet transforms. *IEEE Transactions on Image Processing*, 26(4):1899–1910, April 2017. doi: 10.1109/TIP.2017.2665041.
- J. R. Driscoll and D. M. Healy. Computing Fourier Transforms and Convolutions on the 2-Sphere. *Advances in applied mathematics*, 15(2):202–250, 1994.
- J. Fageot, V. Uhlmann, Zs. Püspöki, B. Beck, M. Unser, and A. Depeursinge. Principled design and implementation of steerable detectors. *arXiv preprint arXiv:1811.00863*, 2018.

- W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(9):891–906, 1991.
- R.A. Gatenby, O. Grove, and R.J. Gillies. Quantitative imaging in cancer evolution and ecology. *Radiology*, 269(1):8–14, 2013.
- K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- S. P. Martin, J. Hofmeister, S. Burgmeister, S. Orso, N. Mili, S. Guerrier, M. P. Victoria-Fesser, P. M. Soccal, F. Triponez, W. Karenovics, N. Mach, A. Depeursinge, C. D. Becker, C. Rampinelli, P. Summers, H. Müller, and X. Montet. Artificial intelligence in lung nodule classification: a radiomics solution. *European Respiratory Journal*, submitted.
- T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, July 2002.
- M. Unser and N. Chenouard. A Unifying Parametric Framework for 2D Steerable Wavelet Transforms. *SIAM Journal on Imaging Sciences*, 6(1):102–135, 2013.
- M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1-2):61–81, 2005. ISSN 0920-5691. doi: 10.1007/s11263-005-4635-4.
- D.A. Varshalovich, A.N. Moskalev, and V.K. Khersonskii. *Quantum theory of angular momentum*. World Scientific, 1988.
- M. Weiler, F.A. Hamprecht, and M. Storath. Learning steerable filters for rotation equivariant CNNs. *arXiv preprint arXiv:1711.07289*, 2017.
- M. Weiler, M. Geiger, M. Welling, W. Boomsma, and T.S. Cohen. 3D steerable CNNs: Learning rotationally equivariant features in volumetric data. *arXiv preprint arXiv:1807.02547*, 2018.
- M. Winkels and T.S. Cohen. 3D G-CNNs for pulmonary nodule detection. *arXiv preprint arXiv:1804.04656*, 2018.
- D.E. Worrall, S. J. Garbin, D. Turmukhambetov, and G.J. Brostow. Harmonic networks: Deep translation and rotation equivariance. *CoRR*, abs/1612.0, 2016. doi: 10.1109/CVPR.2017.758.
- H. Zhang, J. Xue, and K. Dana. Deep TEN: Texture encoding network. *arXiv preprint arXiv:1612.02844*, 2016.
- Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Oriented response networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 519–528, 2017.

## Appendix A. Equivariant Texture Operator

We prove the following result.

**Proposition 1** *A texture operator of the form (1) is equivariant to translations and rotations in the sense of (2).*

**Proof** The equivariance to translations uses that  $(I(\cdot - \mathbf{x}_0) * g)(\mathbf{x}) = (I * g)(\mathbf{x} - \mathbf{x}_0)$ . Applying this to  $g = f(\mathbf{R}\cdot)$ , we deduce

$$\mathcal{G}\{I(\cdot - \mathbf{x}_0)\}(\mathbf{x}) = \max_{\mathbf{R} \in SO(3)} |(I * f(\mathbf{R}\cdot))(\mathbf{x} - \mathbf{x}_0)| = \mathcal{G}\{I\}(\mathbf{x} - \mathbf{x}_0), \quad (7)$$

as expected. For the rotation, we use  $(I(\mathbf{R}_0 \cdot) * g)(\mathbf{x}) = (I * g(\mathbf{R}_0^{-1} \cdot))(\mathbf{R}_0 \mathbf{x})$  applied to  $g = f(\mathbf{R}\cdot)$ , to deduce

$$\mathcal{G}\{I(\mathbf{R}_0 \cdot)\}(\mathbf{x}) = \max_{\mathbf{R} \in SO(3)} |(I * f(\mathbf{R}\mathbf{R}_0^{-1} \cdot))(\mathbf{R}_0 \mathbf{x})| = \max_{\mathbf{R} \in SO(3)} |(I * f(\mathbf{R}\cdot))(\mathbf{R}_0 \mathbf{x})| = \mathcal{G}\{I\}(\mathbf{R}_0 \mathbf{x}), \quad (8)$$

where the second equality simply exploits that  $\mathbf{R}\mathbf{R}_0^{-1}$  describes the complete space  $SO(3)$  of 3D rotations when  $\mathbf{R}$  varies. ■

We remark that the equivariance to translations is simply due to the use of the convolution, while the equivariance to rotations requires the presence of the pooling over 3D rotations in (1).

## Appendix B. Spherical Harmonics

The family of SHs is denoted by  $(Y_{n,m})_{n \geq 0, m \in \{-n, \dots, n\}}$ , where  $n$  is called the degree and  $m$  the order of  $Y_{n,m}$ . SHs form an orthonormal basis for square-integrable functions in the 2D-sphere  $\mathbb{S}^2$ . They are defined as (Driscoll and Healy, 1994)

$$Y_{n,m}(\theta, \phi) = A_{n,m} P_{n,|m|}(\cos(\theta)) e^{im\phi}, \quad (9)$$

with  $A_{n,m} = (-1)^{(m+|m|)/2} \left( \frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!} \right)^{1/2}$  a normalization constant and  $P_{n,|m|}$  the associated Legendre polynomial given for  $0 \leq m \leq n$  by (Abramowitz and Stegun, 1964).

$$P_{n,m}(x) := \frac{(-1)^m}{2^n n!} (1-x^2)^{m/2} \frac{d^{n+m}}{dx^{n+m}} (x^2 - 1)^n. \quad (10)$$

## Appendix C. Real Steerable Filters

A filter  $f$  is real if  $\overline{f(\rho, \theta, \phi)} = f(\rho, \theta, \phi)$  for every  $(\rho, \theta, \phi)$ . For filters given by (3), this means that

$$\overline{h(\rho)} \sum_{n,m} \overline{C_n[m] Y_{n,m}(\theta, \phi)} = h(\rho) \sum_{n,m} C_n[m] Y_{n,m}(\theta, \phi), \quad (11)$$

We use the symmetry of the spherical harmonics,  $\overline{Y_{n,m}} = (-1)^m Y_{n,-m}$ , on the left-hand side and change the sign of  $m$  on the right-hand side to get

$$\sum_{n,m} \overline{h(\rho) C_n[m]} (-1)^m Y_{n,-m}(\theta, \phi) = \sum_{n,m} h(\rho) C_n[-m] Y_{n,-m}(\theta, \phi), \quad (12)$$

The  $Y_{n,m}$  being linearly independent, we deduce that the filter is real if and only if, for any  $\rho, n, m$ ,  $\overline{h(\rho)C_n[m]}(-1)^m = h(\rho)C_n[-m]$ . By imposing that  $h$  is real, i.e.,  $\bar{h} = h$ , we obtain the expected criterion on the  $C_n[m]$  coefficients, which is

$$C_n[-m] = (-1)^m \overline{C_n[m]}, \quad (13)$$

# On the Spatial and Temporal Influence for the Reconstruction of Magnetic Resonance Fingerprinting

**Fabian Balsiger**<sup>1,2,3</sup>

FABIAN.BALSIGER@ARTORG.UNIBE.CH

<sup>1</sup> *Insel Data Science Center, Inselspital, Bern University Hospital, University of Bern, Switzerland*

<sup>2</sup> *NMR Laboratory, Institute of Myology, Neuromuscular Investigation Center, France*

<sup>3</sup> *NMR Laboratory, CEA, DRF, IBFJ, MIRCen, France*

**Olivier Scheidegger**<sup>2,4,5</sup>

OLIVIER.SCHEIDEgger@INSEL.CH

<sup>4</sup> *Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Switzerland*

<sup>5</sup> *Support Center for Advanced Neuroimaging (SCAN), Institute for Diagnostic and Interventional Neuroradiology, Inselspital, Bern University Hospital, University of Bern, Switzerland*

**Pierre G. Carlier**<sup>2,3</sup>

P.CARLIER@INSTITUT-MYOLOGIE.ORG

**Benjamin Marty**<sup>2,3</sup>

B.MARTY@INSTITUT-MYOLOGIE.ORG

**Mauricio Reyes**<sup>1</sup>

MAURICIO.REYES@MED.UNIBE.CH

## Abstract

Magnetic resonance fingerprinting (MRF) is a promising tool for fast and multiparametric quantitative MR imaging. A drawback of MRF, however, is that the reconstruction of the MR maps is computationally demanding and lacks scalability. Several works have been proposed to improve the reconstruction of MRF by deep learning methods. Unfortunately, such methods have never been evaluated on an extensive clinical data set, and there exists no consensus on whether a fingerprint-wise or spatiotemporal reconstruction is favorable. Therefore, we propose a convolutional neural network (CNN) that reconstructs MR maps from MRF-WF, a MRF sequence for neuromuscular diseases. We evaluated the CNN’s performance on a large and highly heterogeneous data set consisting of 95 patients with various neuromuscular diseases. We empirically show the benefit of using the information of neighboring fingerprints and visualize, via occlusion experiments, the importance of temporal frames for the reconstruction.

**Keywords:** Magnetic Resonance Fingerprinting, Image Reconstruction, Convolutional Neural Network, Quantitative Magnetic Resonance Imaging, Neuromuscular Diseases.

## 1. Introduction

Neuromuscular diseases impose a high burden to patients and society regarding severity and economic costs (Larkindale et al., 2014). For such diseases, noninvasive outcome measures to monitor disease progression and treatment effects are desperately needed. A promising tool for obtaining noninvasive outcome measures is magnetic resonance imaging (MRI) (Carlier et al., 2016). However, MRI is rarely performed for outcome measures due to the long acquisition times associated with quantitative MRI, which renders it clinically not applicable. Magnetic resonance fingerprinting (MRF) (Ma et al., 2013) is a novel and promising approach for quantitative MRI, which allows acquiring multiple quantitative MR parameters in one fast scan. The principle of MRF is to pseudo-randomly vary the MR sequence parameters to obtain a unique signal evolution, or fingerprint, per voxel and tissue type. Each fingerprint is then compared to a dictionary of pre-computed fingerprints

to estimate the quantitative MR parameters at interest. For instance, we proposed MRF-WF ([Marty and Carlier, 2018](#)), a MRF sequence for water T1 relaxation time ( $T1_{H2O}$ ) and fat fraction (FF) quantification in skeletal muscles with an acquisition time of only 50 seconds. But while the acquisition time is short, the reconstruction of the  $T1_{H2O}$  and FF maps requires approximately 20 hours per scan using a standard dictionary matching algorithm.

Several methods attempting to accelerate the reconstruction of MRF have been proposed lately. Fast group matching ([Cauley et al., 2015](#)) and search window reduction ([Gómez et al., 2016](#)) aim both at reducing the number of fingerprint comparisons during the dictionary matching. The dictionary matching has also been investigated within the frame of tree data structures and iterative reconstruction ([Cline et al., 2017; Golbabaei et al., 2018b](#)). Further, the compression of the fingerprints has also been investigated ([McGivney et al., 2014; Assländer et al., 2018](#)). However, these approaches suffer mainly by a over-discretized reconstruction, an approximation by compression, and most importantly a poor scalability with more MR parameters. To cope with the mentioned problems, several works have focused on replacing the dictionary matching by learning the mapping between fingerprints and MR parameters by deep learning. The deep learning-based MRF reconstruction can mainly be grouped into fingerprint-wise and spatial-temporal reconstruction. The fingerprint-wise reconstruction takes a single fingerprint as input and estimates the MR parameter(s) at the fingerprint's voxel. The proposed architectures vary from fully-connected networks ([Cohen et al., 2018; Golbabaei et al., 2018a; Barbieri et al., 2018](#)), to fully-convolutional with 1-D convolutions ([Hoppe et al., 2017](#)), the combination of both ([Virtue et al., 2017](#)), and recurrent neural networks ([Oksuz et al., 2019](#)). A limitation of the fingerprint-wise reconstruction is that they do not take advantage of the information between neighboring fingerprints. Recently, spatiotemporal reconstruction by inputting a neighborhood of fingerprints to a neural network has been shown to be beneficial ([Balsiger et al., 2018; Fang et al., 2018](#)). The proposed approaches reconstruct MR parameters from  $5 \times 5$  patches via a convolutional neural network (CNN) ([Balsiger et al., 2018](#)) and MR maps slice-wise with an encoder-decoder CNN architecture ([Fang et al., 2018](#)).

Despite the increasing research being conducted on deep learning-based reconstruction of MRF, there exist several limitations. In general, all studies conducted experiments either on phantom or volunteer data. Furthermore, the data sets were limited in size with the maximum being six scans ([Balsiger et al., 2018](#)). Further, while the studies on spatiotemporal reconstruction have shown a potential advantage over fingerprint-wise reconstruction, it remains unclear to what extent the spatial dimension should be considered in the reconstruction. Last, none of the studies investigated the influence of the temporal frames on the reconstruction. Therefore, we propose a CNN that reconstructs quantitative MR maps from MRF. First, we show its performance on a large and highly heterogeneous data set with 95 scans of patients suffering from various neuromuscular diseases who were imaged at two anatomical regions. Second, we report on three different strategies to incorporate spatial information: i) fingerprint-wise reconstruction ([Cohen et al., 2018](#)), ii) slice-wise reconstruction with spatial pooling operations ([Fang et al., 2018](#)), and iii) our proposed reconstruction with varying sizes of receptive fields. Last, to gain insights into the MR parameter reconstruction, we visualize the importance of the temporal frames via occlusion experiments applied on the temporal domain.

## 2. Materials and Methods

### 2.1. MRF Acquisition and Reconstruction

We acquired the MRF-WF sequence at the legs and thighs levels in patients with various neuro-muscular diseases, resulting in a highly heterogeneous data set with 95 scans (43 female, 52 male; age =  $53.7 \pm 19.5$  years). The MRF-WF acquisition consisted of a non-selective inversion followed by a 1400 radial spokes FLASH echo train (golden angle scheme) with varying echo time (TE), repetition time (TR), and nominal flip angle (FA). We used a field of view of  $350 \times 350 \text{ mm}^2$  with a voxel size of  $1.0 \times 1.0 \times 8.0 \text{ mm}^3$  and five slices per scan, which resulted in a total acquisition time of 50 s. All experiments were performed on a 3 Tesla Siemens MAGNETOM Prisma<sup>fit</sup> scanner (Siemens Healthineers, Erlangen, Germany).

We reconstructed a MRF image space series consisting of 175 temporal frames from the raw  $k$ -space data after the acquisition. The reconstruction involved view sharing with a  $k$ -space weighted image contrast filter with 55 spokes, and compressed sensing with total variation (Marty et al., 2018a,b). This reconstruction resulted in a complex-valued MRF image space series of size  $350 \times 350 \times 175$  voxels per slice. For each scan, T1<sub>H2O</sub>, FF, and transmit field efficacy (B1) reference MR maps were reconstructed from the MRF image space series using dictionary matching (Marty and Carlier, 2018). The total reconstruction time was approximately 20 hours per scan using a standard desktop computer (2.6 GHz Intel Xenon E5-2630, 48 GB memory).

### 2.2. CNN Map Reconstruction

We propose a CNN architecture that reconstructs MR maps patch-wise from MRF image space series. Let us consider a 2-D+time MRF image space series  $I \in \mathbb{C}^{X \times Y \times T}$  and its corresponding reference MR maps  $Q \in \mathbb{R}^{X \times Y \times M}$  with  $X \times Y = 350 \times 350$  being the spatial dimension,  $T = 175$  being the number of temporal frames, and  $M = 3$  being the number of MR maps (T1<sub>H2O</sub>, FF, B1). Our CNN learns the mapping  $\mathcal{M} : I_P \rightarrow Q_P$ , i.e. to reconstruct a patch  $Q_P \in \mathbb{R}^{QP_X \times QP_Y \times M} \subset Q$  of the MR maps from a patch  $I_P \in \mathbb{C}^{IP_X \times IP_Y \times T} \subset I$  of the 2-D+time MRF image space series. In our experiments, we reconstructed non-overlapping output patches of a size  $QP_X \times QP_Y = 32 \times 32$  with the input patch size being set to  $IP_X \times IP_Y = 42 \times 42$  according to the spatial receptive field of the CNN ( $11 \times 11$ ).<sup>1</sup>

#### 2.2.1. PRE-PROCESSING

We normalized the real and imaginary parts of each MRF image space series  $I$  to zero mean and unit variance. Each MR map has been normalized to the range  $[0, 1]$  using the minimum and maximum values of the entire data set.

#### 2.2.2. CNN ARCHITECTURE

Our CNN architecture consists of two types of blocks: a temporal block that leverages the temporal information of the fingerprints, and a spatial block that leverages the spatial correlation between neighboring fingerprints. After each temporal block follows a spatial block, with this sequence being repeated five times. We treat the temporal dimension  $T$  as feature channels, and therefore,

---

1. The size of  $Q_P$  is mainly restricted by GPU memory. We could not determine a statistically significant difference to other patch sizes.

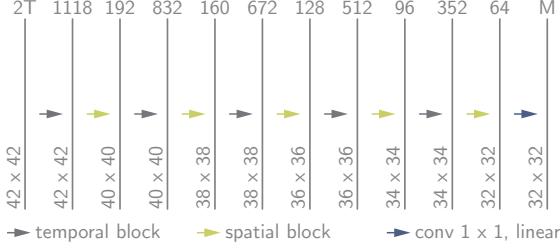


Figure 1: The proposed CNN architecture. The numbers at the top of the bars denote the number of feature channels ( $T = 175$  and  $M = 3$  for MRF-WF), and numbers at the lower left of the bars denote the spatial patch size.

the concatenated real and imaginary parts of the complex-valued  $I_P$  serve as input, i.e., we used a real-valued input  $I_{P\text{real}} \in \mathbb{R}^{IP_X \times IP_Y \times 2T}$ . The output  $Q_P$  is reconstructed after a final 2-D convolution with kernel size of  $1 \times 1$ ,  $M$  feature channels, and a linear activation (Figure 1).

Our temporal blocks learn features along the fingerprints by applying  $1 \times 1$  convolutions and are inspired by dense blocks (Huang et al., 2017). The temporal blocks contain four layers, and each layer is connected to the output feature channels of every preceding layer. The number of feature channels, or growth rate,  $n$  is equal for the four layers in a temporal block, and decreases within the CNN from 192, 160, 128, 92, to 64. In order to reuse features and to further facilitate the gradient flow, we additionally concatenate the input of a temporal block to its output arriving at  $4 \cdot n + n_{in}$  features channels after each temporal block, with  $n_{in}$  being the number of input feature channels to the temporal block. Each layer in our temporal block consists of a 2-D convolution with kernel size of  $1 \times 1$ , rectified linear unit (ReLU) activation function (Glorot et al., 2011), and batch normalization (Ioffe and Szegedy, 2015). After each temporal block, our spatial blocks extract spatial features by performing a 2-D convolution with a kernel size of  $3 \times 3$  (valid padding) followed by a ReLU activation function, and batch normalization. The number of feature channels of the spatial blocks decreases equally to the growth rate of the temporal blocks within the CNN. Code is available online<sup>2</sup>.

### 2.2.3. TRAINING

We trained our CNN for 100 epochs with a batch size of 50 randomly selected patches. We used an Adam optimizer (Kingma and Ba, 2015) to minimize a mean squared error loss with a learning rate of 0.001,  $\beta_1 = 0.9$ , and  $\beta_2 = 0.999$ . The CNN was implemented with TensorFlow 1.10.0 (Google, Mountain View, CA, U.S.) and Python 3.6.5 (Python Software Foundation, Wilmington, DE, U.S.).

## 2.3. Comparison and Evaluation

We compared our CNN to three other deep learning-based MRF reconstruction methods. The first method is a neural network with two hidden fully-connected layers (Cohen et al., 2018). The method

2. <https://github.com/fabianbalsiger/mrf-reconstruction-midl2019>

requires single fingerprint as input and is, therefore, an extreme variant where no neighboring fingerprint information is considered. The second method is an encoder-decoder CNN with two spatial pooling operations (Fang et al., 2018). The method reconstructs entire MR map slices and hence represents the opposite extreme variant due to the spatial pooling operations, which increase the receptive field drastically compared to our method. We implemented both methods as originally described with slight modifications for our MRF sequence and we used the same pre-processing and training scheme as for our method. Last, we also show the reconstructions of our previous architecture with a receptive field of  $5 \times 5$  (Balsiger et al., 2018).

We evaluated the performance of our CNN and the two compared methods by randomly splitting our data set into training/validation/testing sets ( $n=55/20/20$ ). Note that we purposely did not apply any stratification regarding neuromuscular disease or anatomical region to evaluate the robustness to the heterogeneous data set. The best performing model on the validation set, in terms of the normalized root mean squared error (NRMSE), was selected and applied to the testing set. We calculated the NRMSE, peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM) (Wang et al., 2004) between the predicted and the reference maps on a scan level. Additionally, we calculated the coefficient of determination ( $R^2$ ) between the predicted and reference mean values of manually segmented region of interests within the major muscles across all scans. For the calculations of the NRMSE and PSNR, background voxels were excluded using a mask segmented by thresholding an anatomical image (out-of-phase). The mask for the  $T1_{H2O}$  map was further processed to exclude voxels with a FF higher than 0.65 because the confidence of  $T1_{H2O}$  values is highly decreased at high FFs. For the SSIM, we used a window size of  $7 \times 7$ ,  $K_1 = 0.01$ ,  $K_2 = 0.03$ , and  $L$  was set to the maximum value of the reference map.

### 3. Experiments and Results

We first present the comparison between our method and (Cohen et al., 2018; Fang et al., 2018; Balsiger et al., 2018) for the deep learning-based reconstruction of MRF-WF. Second, we investigate the influence of the CNN’s spatial receptive field on the reconstruction. Third, we visualize the importance of the temporal frames on the reconstruction.

#### 3.1. Deep Learning-based Reconstruction of MRF-WF

Figure 2 shows the reconstruction of  $T1_{H2O}$ , FF, and B1 maps of a representative case. Compared to our method, the fingerprint-wise method of (Cohen et al., 2018) produced noisier and the slice-wise with spatial pooling operations method of (Fang et al., 2018) smoothed map reconstructions. Further, we are able to improve over our previous architecture (Balsiger et al., 2018). Quantitatively, our method yielded the best reconstructed MR maps for all evaluation metrics (Table 1).

#### 3.2. Influence of the Receptive Field

To this date, it is unclear how the information of neighboring fingerprints affects the deep learning-based MR map reconstruction. We think that an optimal solution lies between the extreme variants of fingerprint-wise reconstruction and reconstruction including spatial pooling operations with large receptive fields. Therefore, we relied on a patch-wise MR map reconstruction and determined the effect of the receptive field by experimentation. We modified the spatial blocks in our CNN by performing convolutions with a kernel size of  $1 \times 1$  instead of  $3 \times 3$  to obtain CNNs with receptive

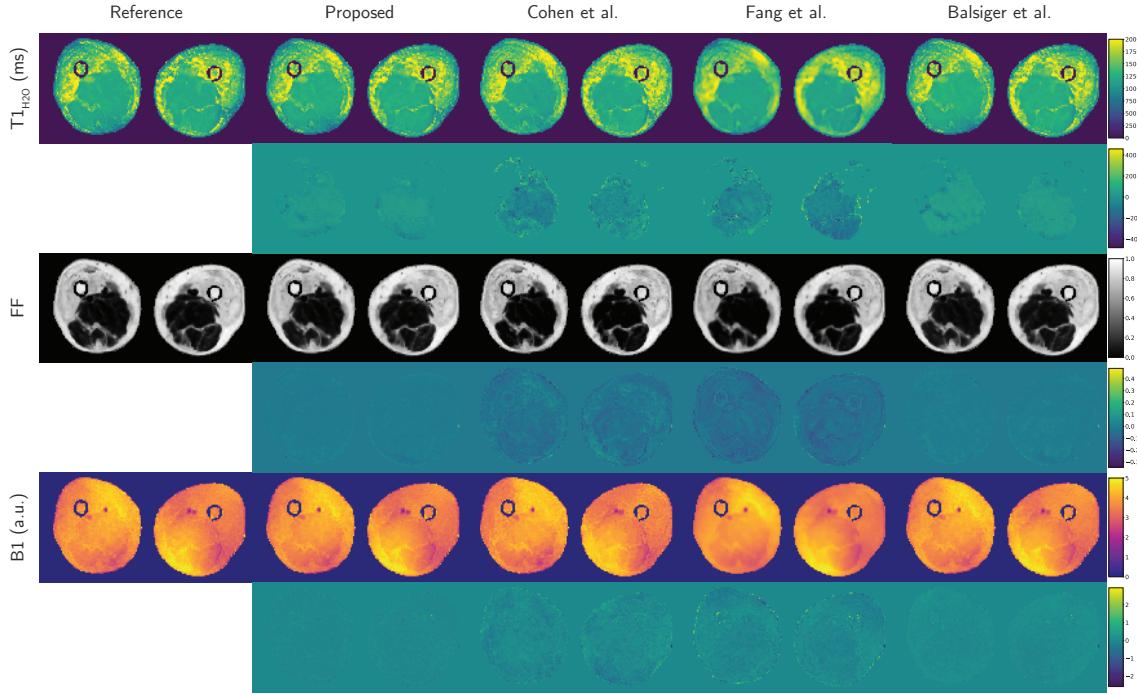


Figure 2:  $T1_{H2O}$ , FF, and B1 maps of a 73 years old male patient with inclusion body myositis. The reference maps, the reconstructions, and the reconstruction errors (reference minus reconstruction) of our and the compared methods are shown. a.u.: arbitrary unit.

Table 1: Results for the MR map reconstruction.

Metric	MR map	Proposed	Cohen et al.	Fang et al.	Balsiger et al.
NRMSE	$T1_{H2O}$	<b><math>0.024 \pm 0.016</math></b>	$0.065 \pm 0.026$	$0.064 \pm 0.028$	$0.032 \pm 0.017$
	FF	<b><math>0.017 \pm 0.008</math></b>	$0.042 \pm 0.005$	$0.056 \pm 0.005$	$0.021 \pm 0.007$
	B1	<b><math>0.024 \pm 0.007</math></b>	$0.051 \pm 0.006$	$0.062 \pm 0.006$	$0.035 \pm 0.006$
PSNR	$T1_{H2O}$	<b><math>36.16 \pm 4.211</math></b>	$26.53 \pm 1.869$	$26.85 \pm 2.557$	$33.27 \pm 3.485$
	FF	<b><math>35.84 \pm 3.038</math></b>	$27.50 \pm 0.969$	$25.00 \pm 0.799$	$33.90 \pm 2.564$
	B1	<b><math>34.27 \pm 1.856</math></b>	$27.59 \pm 0.864$	$25.88 \pm 0.842$	$31.08 \pm 1.135$
SSIM	$T1_{H2O}$	<b><math>0.978 \pm 0.011</math></b>	$0.931 \pm 0.029$	$0.915 \pm 0.040$	$0.972 \pm 0.013$
	FF	<b><math>0.990 \pm 0.007</math></b>	$0.955 \pm 0.024$	$0.953 \pm 0.033$	$0.985 \pm 0.012$
	B1	<b><math>0.977 \pm 0.009</math></b>	$0.934 \pm 0.025$	$0.926 \pm 0.030$	$0.974 \pm 0.008$
R2	$T1_{H2O}$	<b><math>0.964</math></b>	$0.888$	$0.909$	$0.949$
	FF	<b><math>0.999</math></b>	$0.998$	$0.996$	<b><math>0.999</math></b>
	B1	<b><math>0.998</math></b>	$0.967$	$0.950$	$0.993$

fields lower than  $11 \times 11$ , and added additional  $3 \times 3$  convolutions with 64 channels before the last  $1 \times 1$  convolutional layer of our CNN to obtain CNNs with receptive fields up to  $21 \times 21$ .<sup>3</sup> Figure 3 shows the NRMSE for the  $T1_{H2O}$  map reconstruction, which is the most difficult, with varying receptive fields of our CNN. The optimal receptive field for MRF-WF is  $11 \times 11$ , with a statistically significant better reconstruction compared to receptive fields of size  $5 \times 5$  and lower. Figures for the FF and B1 map reconstruction can be found in the Appendix A.

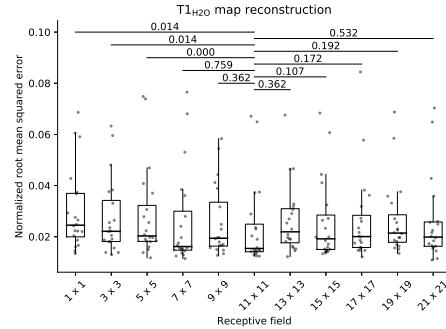


Figure 3: Effect of the receptive field on the  $T1_{H2O}$  map reconstruction. Paired Wilcoxon signed rank test with significance level of 0.05 and Bonferroni correction for multiple comparisons.

### 3.3. Importance of the Temporal Frames

We adopted the occlusion experiments proposed by (Zeiler and Fergus, 2014) for MRF to examine the importance of the temporal frames of the fingerprint for the MR map reconstruction. We applied the occlusion in the temporal domain, which allows for simple yet effective visualization of the importance of the temporal frames. Therefore, we blacked out each  $t$ -th temporal frame and reconstructed the MR maps with this occluded MRF image space series. We then considered the absolute difference of the NRMSE to the non-occluded MRF image space series as the importance of the  $t$ -th temporal image for the reconstruction. The importance of each temporal frame  $t$  is shown in Figure 4a. We observe that the first temporal frames, after the non-selective inversion pulse, have a high importance for the  $T1_{H2O}$  and FF map reconstruction. Additionally, the importance correlates with changes of the MRF-WF sequence parameters over the course of the acquisition (cf. Figure 4b). Note that we could also observe the same pattern for the importance when running the occlusion experiment on modified versions of our network.

## 4. Discussion

We evaluated a CNN consisting of temporal and spatial blocks for the MR map reconstruction from MRF-WF, and compared it to the current state-of-the-art deep learning-based MRF reconstruction

3. We remark that this results in a slightly different number of learnable parameters, which could be compensated. However, we decided not to do so as this would change the CNN architecture.

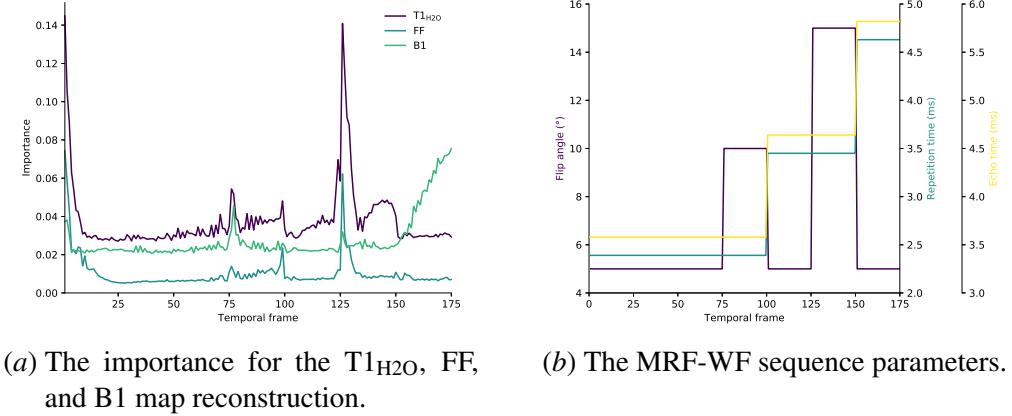


Figure 4: The importance of the temporal frames for the MR map reconstruction.

on a large and highly heterogeneous data set consisting only of patient scans. Our proposed CNN outperformed the compared methods quantitatively and qualitatively. We attribute the good performance of our CNN to the trade-off between temporal and spatial feature learning. An indication for the superiority of such a spatiotemporal trade-off is also the noisier or over-smoothed reconstructions of the compared methods.

We conducted an experiment to show the influence of the spatial receptive field on the reconstruction. It is evident that the reconstruction benefits from the information of neighboring fingerprints, however, only to some extent. For MRF-WF, a receptive field larger than  $11 \times 11$  voxels is not beneficial, which we also attribute to the highly diverse appearance of neuromuscular pathologies. Especially incorporating spatial features from multiple scales as with the method of (Fang et al., 2018) yielded worse reconstructions. For future developments, it is therefore essential to consider the application of the MRF sequence. MRF scans with small or highly diverse pathologies might benefit from the information of neighboring fingerprints but only to a certain extent. We also note that the optimal receptive field is certainly depended on the in-plane voxel size and the  $k$ -space sampling.

Visualizing the importance of the temporal frames might be a useful tool for further developments of MRF reconstruction, and the MRF sequence development itself. From a MR physical point of view, the first temporal frames should be sensitive to  $T1_{H2O}$  due to the applied inversion pulse at the beginning of the MRF-WF acquisition. We could observe a high importance of the first few temporal frames for the  $T1_{H2O}$  but also for the FF reconstruction by our occlusion experiment. Further, the high correlation of the importance to changes in the MRF sequence parameters might suggest using a smoother transition of parameter values. However, further work needs to be done to better understand the reconstruction. For instance, it remains unclear from a MR physical point of view, why the last 25 temporal frames are of high importance for the B1 map reconstruction.

## 5. Conclusion

We proposed a spatiotemporal CNN to reconstruct MR maps from MRF, and showed empirically that a deep learning-based reconstruction benefits from the information between neighboring fingerprints and learns features that can make sense from a MR physical point of view.

## Acknowledgments

This research was partially supported by the Swiss National Science Foundation (SNSF) under grant number 184273. The authors thank the NVIDIA corporation for the donation of a GPU and appreciate the valuable discussions with Pierre-Yves Baudin and Alain Jungo.

## References

- Jakob Assländer, Martijn A. Cloos, Florian Knoll, Daniel K. Sodickson, Jürgen Hennig, and Riccardo Lattanzi. Low Rank Alternating Direction Method of Multipliers Reconstruction for MR Fingerprinting. *Magnetic Resonance in Medicine*, 79(1):83–96, 2018. ISSN 07403194. doi: 10.1002/mrm.26639.
- Fabian Balsiger, Amaresha Shridhar Konar, Shivaprasad Chikop, Vimal Chandran, Olivier Scheidegger, Sairam Geethanath, and Mauricio Reyes. Magnetic Resonance Fingerprinting Reconstruction via Spatiotemporal Convolutional Neural Networks. In Florian Knoll, Andreas Maier, and Daniel Rueckert, editors, *Machine Learning for Medical Image Reconstruction*, volume 11074 of *Lecture Notes in Computer Science*, pages 39–46. Springer, Cham, 2018. doi: 10.1007/978-3-030-00129-2\_5.
- Marco Barbieri, Leonardo Brizi, Enrico Giampieri, Francesco Solera, Gastone Castellani, Claudia Testa, and Daniel Remondini. Circumventing the Curse of Dimensionality in Magnetic Resonance Fingerprinting through a Deep Learning Approach. *arXiv preprint arXiv:1811.11477*, 2018.
- Pierre G Carlier, Benjamin Marty, Olivier Scheidegger, Paulo Loureiro de Sousa, Pierre-Yves Baudin, Eduard Snezhko, and Dmitry Vlodavets. Skeletal Muscle Quantitative Nuclear Magnetic Resonance Imaging and Spectroscopy as an Outcome Measure for Clinical Trials. *Journal of Neuromuscular Diseases*, 3(1):1–28, 2016. ISSN 2214-3599. doi: 10.3233/JND-160145.
- Stephen F. Cauley, Kawin Setsompop, Dan Ma, Yun Jiang, Huihui Ye, Elfar Adalsteinsson, Mark A. Griswold, and Lawrence L. Wald. Fast Group Matching for MR Fingerprinting Reconstruction. *Magnetic Resonance in Medicine*, 74(2):523–528, 2015. ISSN 07403194. doi: 10.1002/mrm.25439.
- Christopher C. Cline, Xiao Chen, Boris Mailhe, Qiu Wang, Josef Pfeuffer, Mathias Nittka, Mark A. Griswold, Peter Speier, and Mariappan S. Nadar. AIR-MRF: Accelerated iterative reconstruction for magnetic resonance fingerprinting. *Magnetic Resonance Imaging*, 41:29–40, 2017. ISSN 0730-725X. doi: 10.1016/J.MRI.2017.07.007.
- Ouri Cohen, Bo Zhu, and Matthew S. Rosen. MR fingerprinting Deep RecOnstruction NEtwork (DRONE). *Magnetic Resonance in Medicine*, 80(3):885–894, 2018. ISSN 07403194. doi: 10.1002/mrm.27198.

Zhenghan Fang, Yong Chen, Mingxia Liu, Yiqiang Zhan, Weili Lin, and Dinggang Shen. Deep Learning for Fast and Spatially-Constrained Tissue Quantification from Highly-Undersampled Data in Magnetic Resonance Fingerprinting (MRF). In Yinghuan Shi, Heung-Il Suk, and Mingxia Liu, editors, *Machine Learning in Medical Imaging*, volume 11046 of *Lecture Notes in Computer Science*, pages 398–405. Springer, Cham, 2018. doi: 10.1007/978-3-030-00919-9\_46.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep Sparse Rectifier Neural Networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, 2011. PMLR.

Mohammad Golbabae, Dongdong Chen, Pedro A. Gómez, Marion I. Menzel, and Mike E. Davies. Geometry of Deep Learning for Magnetic Resonance Fingerprinting. *arXiv preprint arXiv:1809.01749*, 2018a. URL <http://arxiv.org/abs/1809.01749>.

Mohammad Golbabae, Zhouye Chen, Yves Wiaux, and Mike Davies. CoverBLIP: accelerated and scalable iterative matched-filtering for Magnetic Resonance Fingerprint reconstruction. *arXiv preprint arXiv:1810.01967*, 2018b.

Pedro A. Gómez, Miguel Molina-Romero, Cagdas Ulas, Guido Bounincontri, Jonathan I. Sperl, Derek K. Jones, Marion I. Menzel, and Bjoern H. Menze. Simultaneous Parameter Mapping, Modality Synthesis, and Anatomical Labeling of the Brain with MR Fingerprinting. In Sébastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 579–586, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46726-9. doi: 10.1007/978-3-319-46726-9\_67.

Elisabeth Hoppe, Gregor Körzdörfer, Tobias Würfl, Jens Wetzl, Felix Lugauer, Josef Pfeuffer, and Andreas Maier. Deep Learning for Magnetic Resonance Fingerprinting: A New Approach for Predicting Quantitative Parameter Values from Time Series. In R. Röhrlig, A. Timmer, H. Binder, and U. Sax, editors, *German Medical Data Sciences: Visions and Bridges*, volume 243, pages 202–206, Oldenburg, 2017. Oldenburg. doi: 10.3233/978-1-61499-808-2-202.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269. IEEE, 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.243.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 448–456, Lille, 2015. PMLR.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations 2015*, pages 1–15, 2015. ISBN 9781450300728. doi: 10.1145/1830483.1830503.

Jane Larkindale, Wenya Yang, Paul F. Hogan, Carol J. Simon, Yiduo Zhang, Anjali Jain, Elizabeth M. Habeeb-Louks, Annie Kennedy, and Valerie A. Cwik. Cost of illness for neuromuscular

diseases in the United States. *Muscle & Nerve*, 49(3):431–438, 2014. ISSN 0148639X. doi: 10.1002/mus.23942.

Dan Ma, Vikas Gulani, Nicole Seiberlich, Kecheng Liu, Jeffrey L Sunshine, Jeffrey L Duerk, and Mark A Griswold. Magnetic resonance fingerprinting. *Nature*, 495(7440):187–192, 2013. ISSN 0028-0836. doi: 10.1038/nature11971.

Benjamin Marty and Pierre G Carlier. Quantification of water T1 and fat fraction in skeletal muscle tissue using an optimal MR fingerprinting radial sequence (MRF-WF). In *International Society for Magnetic Resonance in Medicine*, 2018.

Benjamin Marty, B. Coppa, and Pierre G. Carlier. Fast, Precise, and Accurate Myocardial T1 Mapping Using a Radial MOLLI Sequence With FLASH Readout. *Magnetic Resonance in Medicine*, 79(3):1387–1398, 2018a. ISSN 07403194. doi: 10.1002/mrm.26795.

Benjamin Marty, Bertrand Coppa, and Pierre G. Carlier. Monitoring skeletal muscle chronic fatty degenerations with fast T1-mapping. *European Radiology*, 28(11):4662–4668, 2018b. ISSN 0938-7994. doi: 10.1007/s00330-018-5433-z.

Debra F. McGivney, Eric Pierre, Dan Ma, Yun Jiang, Haris Saybasili, Vikas Gulani, and Mark A. Griswold. SVD Compression for Magnetic Resonance Fingerprinting in the Time Domain. *IEEE Transactions on Medical Imaging*, 33(12):2311–2322, 2014. ISSN 0278-0062. doi: 10.1109/TMI.2014.2337321.

Ilkay Oksuz, Gastao Cruz, James Clough, Aurelien Bustin, Nicolo Fuin, Rene M. Botnar, Claudia Prieto, Andrew P. King, and Julia A. Schnabel. Magnetic Resonance Fingerprinting using Recurrent Neural Networks. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019. URL <http://arxiv.org/abs/1812.08155>.

Patrick Virtue, Stella X. Yu, and Michael Lustig. Better than real: Complex-valued neural nets for MRI fingerprinting. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3953–3957. IEEE, 2017. ISBN 978-1-5090-2175-8. doi: 10.1109/ICIP.2017.8297024.

Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. ISSN 1057-7149. doi: 10.1109/TIP.2003.819861.

Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer International Publishing, Cham, 2014. doi: 10.1007/978-3-319-10590-1\_53.

## Appendix A. Influence of the Receptive Field on the FF and B1 Map Reconstruction

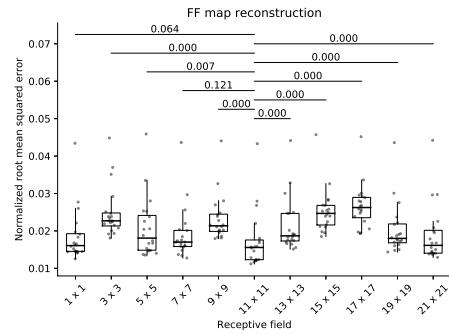


Figure 5: Effect of the receptive field on the FF map reconstruction. Paired Wilcoxon signed rank test with significance level of 0.05 and Bonferroni correction for multiple comparisons.

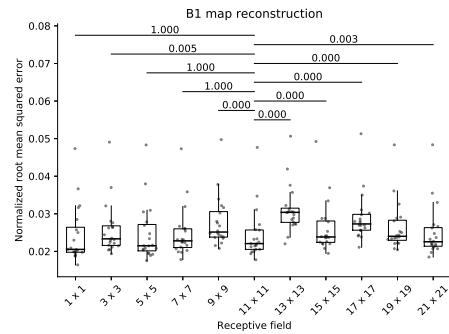


Figure 6: Effect of the receptive field on the B1 map reconstruction. Paired Wilcoxon signed rank test with significance level of 0.05 and Bonferroni correction for multiple comparisons.

# Image Synthesis with a Convolutional Capsule Generative Adversarial Network

**Cher Bass**<sup>1,2,3</sup>

C.BASS14@IMPERIAL.AC.UK

**Tianhong Dai**<sup>1</sup>

TIANHONG.DAI15@IMPERIAL.AC.UK

**Benjamin Billot**<sup>1</sup>

BENJAMIN.BILLOT.18@UCL.AC.UK

**Kai Arulkumaran**<sup>1</sup>

KAILASH.ARULKUMARAN13@IMPERIAL.AC.UK

**Antonia Creswell**<sup>1</sup>

ANTONIA.CRESWELL11@IMPERIAL.AC.UK

**Claudia Clopath**<sup>1</sup>

C.CLOPATH@IMPERIAL.AC.UK

**Vincenzo De Paola**<sup>3</sup>

VINCENZO.DEPAOLA@CSC.MRC.AC.UK

**Anil Anthony Bharath**<sup>1</sup>

A.BHARATH@IMPERIAL.AC.UK

<sup>1</sup> Department of Bioengineering, Imperial College London, UK

<sup>2</sup> Centre for Neurotechnology, Imperial College London, UK

<sup>3</sup> MRC Clinical Science Centre, Faculty of Medicine, Imperial College London, London, UK

## Abstract

Machine learning for biomedical imaging often suffers from a lack of labelled training data. One solution is to use generative models to synthesise more data. To this end, we introduce CapsPix2Pix, which combines convolutional capsules with the pix2pix framework, to synthesise images conditioned on class segmentation labels. We apply our approach to a new biomedical dataset of cortical axons imaged by two-photon microscopy, as a method of data augmentation for small datasets. We evaluate performance both qualitatively and quantitatively. Quantitative evaluation is performed by using image data generated by either CapsPix2Pix or pix2pix to train a U-net on a segmentation task, then testing on real microscopy data. Our method quantitatively performs as well as pix2pix, with an order of magnitude fewer parameters. Additionally, CapsPix2Pix is far more capable at synthesising images of different appearance, but the same underlying geometry. Finally, qualitative analysis of the features learned by CapsPix2Pix suggests that individual capsules capture diverse and often semantically meaningful groups of features, covering structures such as synapses, axons and noise.

**Keywords:** Capsule Network, Generative Adversarial Network, Neurons, Axons, Synthetic Data, Segmentation, Image Synthesis, Image-to-Image Translation

## 1. Introduction

Deep neural networks (DNNs) have significantly advanced the state-of-the-art in biomedical image analysis, particularly where data is well curated for specific tasks such as segmentation and classification (Ronneberger et al., 2015; Frid-Adar et al., 2018b). However, there remain significant challenges. A key problem in analysing biomedical datasets is inadequate quantities of data for training. This may occur where data is scarce, or the expert resources are needed for curated ground truth. We may specifically care about learning from a few data points, as animal or patient data is often restricted.

Recent years have shown that synthetic data, in combination with or even without real data, can be used to effectively train computer vision systems (Gaidon et al., 2018). One such approach would be to train a generative model on modest amounts of labelled data, and use this to augment the dataset with synthesised images. To achieve this, we introduce a convolutional capsule generative adversarial network (GAN), CapsPix2Pix (Figure 1), to synthesise images conditioned on segmentation labels. Through experimental evaluation we show that our method quantitatively and qualitatively matches or outperforms pix2pix, a state-of-the-art conditional image synthesis model (Isola et al., 2017).

In particular, we use our method to synthesise images based on a new 152-image cortical axon dataset, captured with a two-photon microscope in the mouse cortex. We use the synthesised images to pretrain a segmentation model, and show that it improves performance over using only the original dataset. Better segmentation performance allows us to better automate the study of neurons. This is relevant for the field of experimental neuroscience, where there is interest in examining the structure of neurons under different conditions, or in response to a stimuli. We believe that our results validate the use of CapsPix2Pix, which could be applied to other biomedical datasets and downstream tasks.

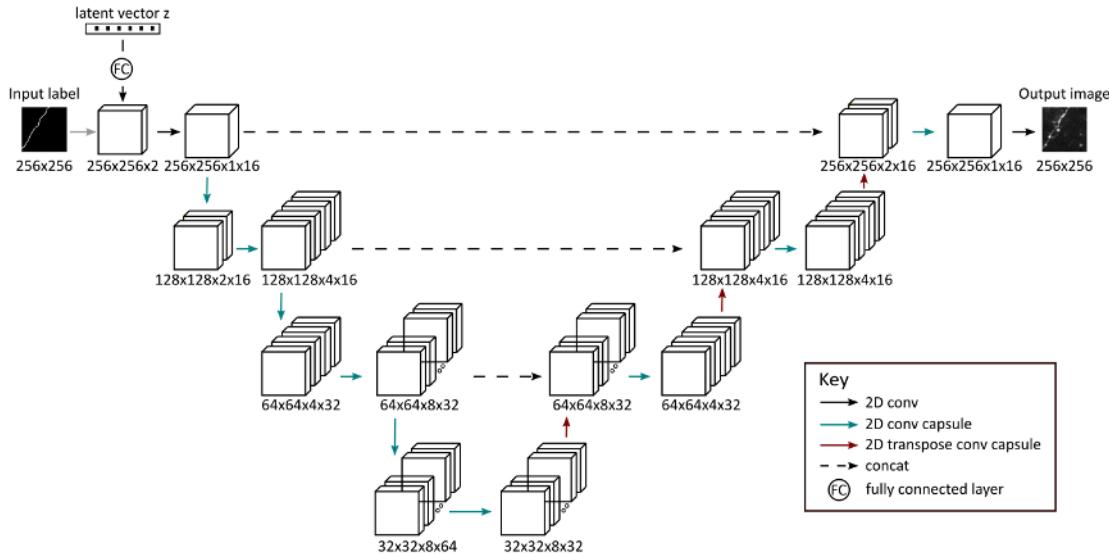


Figure 1: CapsPix2Pix generator architecture.

### 1.1. Summary of Our Contributions

We introduce for the first time a convolutional capsule network in the GAN framework. We propose a convolutional capsule architecture—CapsPix2Pix—for conditional image generation, that utilises an input label and a latent vector (Figure 1). We show that CapsPix2Pix quantitatively performs as well as the state-of-the-art pix2pix (Isola et al., 2017) for generating realistic images with our dataset, while drastically reducing the number of network parameters: 7 $\times$  smaller than pix2pix (7.9M vs. 50M parameters). We show that capsules capture qualitatively different features for

the relevant structures in our dataset, and can create varied synthetic images from the same labels (Figures A5 and 2). We also explore the features learned by convolutional capsules. Through this we find that they group similar features in the same capsule. Finally, we present a new dataset of cortical neurons and segmentation labels collected using two-photon microscopy in the mouse cortex. The dataset is available at <https://doi.org/10.5281/zenodo.2559237>, and the code for our method can be found at <https://github.com/CheBass/CapsPix2Pix>.

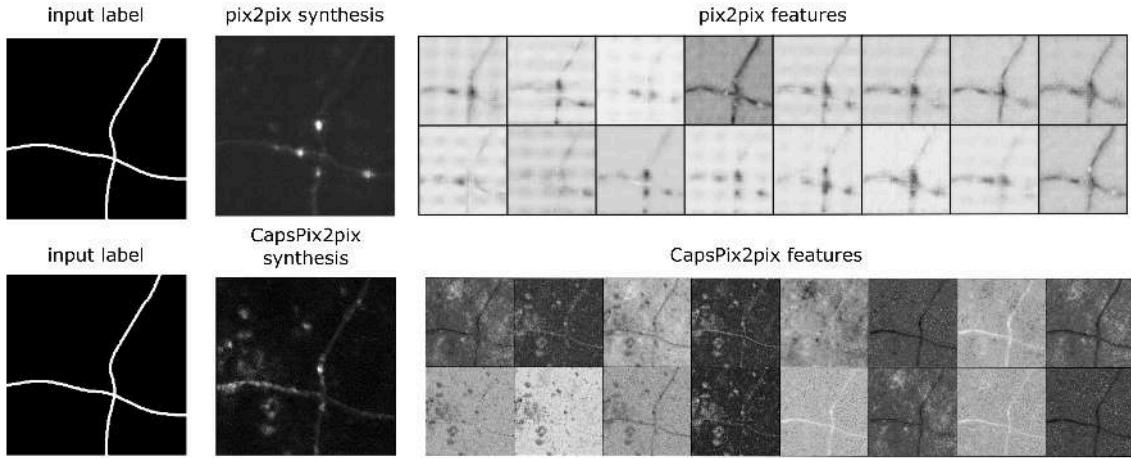


Figure 2: Comparison of the features of pix2pix (16/64 final layer activations— see Figure A7 for all activations) and CapsPix2Pix (16/16 final layer activations) from the same geometric description of an axon.

## 2. Background and Related Work

Recent years have seen the introduction of several methods that are capable of conditional image synthesis. Such methods may take as input segmentation labels (Isola et al., 2017), object bounding boxes (Reed et al., 2016b) or even text (Reed et al., 2016a), and produce realistic images conditioned on this information. For biomedical data, the ability to synthesise new samples would allow us to apply powerful supervised learning methods to datasets with few labels—effectively, semi-supervised learning by synthesising new labelled samples. Accordingly, some prior work has used such methods to generate synthetic biomedical datasets (Hinterstoisser et al., 2017; Alzantot et al., 2017; Frid-Adar et al., 2018a,b; Sixt et al., 2018; Korkinof et al., 2018; Baur et al., 2018).

In this work we build upon pix2pix (Isola et al., 2017), a conditional GAN (cGAN) (Goodfellow et al., 2014; Mirza and Osindero, 2014) that is capable of synthesising images conditional on segmentation labels. We augment pix2pix with convolutional capsules (Sabour et al., 2017; LaLonde and Bagci, 2018), which are proposed to be better able to capture relationships between features than standard convolutional NNs (CNNs).

## 2.1. Generative Adversarial Networks

GANs are a type of implicit generative model that map latent vectors  $z \sim P_z(z)$  to samples  $y$ , via a generator model  $G: z \rightarrow y$  (Goodfellow et al., 2014). They can be extended to the conditional setting (Mirza and Osindero, 2014), where the mapping is from  $z$  and an additional input label  $x$ , such that  $G: \{x, z\} \rightarrow y$ . For the case of our biomedical dataset,  $x$  represents the geometry of the structures to be synthesised,  $z$  captures additional properties of the data distribution, such as imaging noise and variations in axon intensities, and  $y$  is a sample resembling a cortical axon image.

GAN training also involves a discriminator model  $D: \{x, y\} \rightarrow [0, 1]$ , where  $D$  is shown both real and synthetic image-label pairs and is trained to distinguish between them via a binary classification task.  $G$  is then trained to generate samples that fool the discriminator. Training is formulated as a two-player minimax game, where the objective is to find a Nash equilibrium for both models:

$$\min_G \max_D L_{cGAN}(G, D) = \mathbb{E}_{x, y \sim P_{\text{data}}(x, y)} [\log D(x, y)] + \mathbb{E}_{x \sim P_{\text{data}}(x), z \sim P_z(z)} [\log (1 - D(x, G(x, z)))] \quad (1)$$

For more information on (conditional and unconditional) GANs we refer readers to Goodfellow (2016); Creswell et al. (2018).

The pix2pix network (Isola et al., 2017) is a fully-convolutional cGAN based on a U-net-style encoder-decoder architecture (Ronneberger et al., 2015). In order to encourage the final output to better adhere to the structure of the label, they also minimise an  $L_1$  loss between synthesised and real images with the corresponding label:

$$\min_G L_1(G) = \mathbb{E}_{x, y \sim P_{\text{data}}(x, y), z \sim P_z(z)} [\|y - G(x, z)\|_1] \quad (2)$$

The final objective, which we also use for CapsPix2Pix, is the value function  $V(G, D)$ :

$$\min_G \max_D V(G, D) = L_{cGAN}(G, D) + \lambda L_1(G) \quad (3)$$

where  $\lambda = 0.1$  for pix2pix, and  $\lambda = 1$  for CapsPix2Pix (see Appendix C for results comparing  $\lambda \in \{0.1, 1\}$  for CapsPix2Pix). The original formulation of pix2pix also forgoes using  $z \sim P_z(z)$  to generate a distribution of images, and instead uses dropout to stochastically generate outputs, as they found that their generator's outputs were largely independent of  $z$ . We instead retain the approach of inputting random latent vectors to drive CapsPix2Pix, as we found that it was capable of learning meaningful manifold in latent space (Figure 3).

## 2.2. Capsule Networks

Capsule networks were first introduced by Sabour et al. (2017), and were made to better encode spatial relationships between features than standard CNNs. As opposed to standard DNNs, where scalar values represent a feature, capsules output vectors, where the orientation of the vector represents properties (e.g., pose, texture, etc.), and the magnitude represents the probability of the feature being present. The second key component of capsule networks is “dynamic routing”, an iterative algorithm in which outputs of capsules are routed to capsules in the layer above based on how well their predictions agree. We detail the dynamic routing algorithm in Algorithm 1.

However, traditional capsules use a high-dimensional transformation matrix, and thus were only applied to small images. The size of the weight matrix is also fixed based on the input size, and so

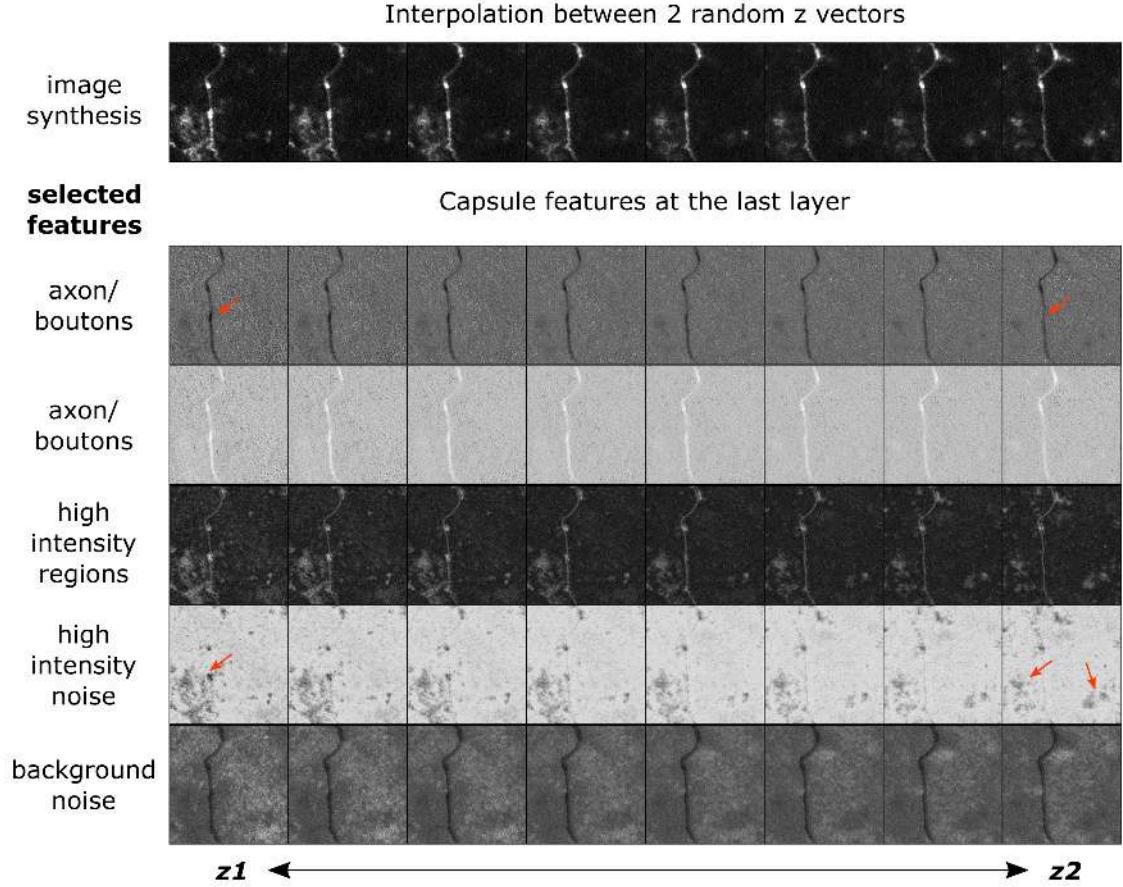


Figure 3: Linear interpolation between two  $z$  vectors for CapsPix2Pix. We show 5 selected features (activations) from our 16D capsule at the last layer. The red arrows in the axon/ boutons row, point to an example of how a bouton appears and disappears in the feature space, when interpolating the  $z$  vectors. The red arrows in the high intensity noise row, point to changing high intensity noise.

the same network cannot be applied to different sized images. [LaLonde and Bagci \(2018\)](#) solved this issue by introducing convolutional capsules, and successfully applied them to a segmentation task with large images ( $512 \times 512$  pixels). As in standard CNNs, convolutional capsules benefit from a smaller memory usage and faster computation. By using a primarily fully-convolutional structure, we are also able to analyse intermediate layer activations (see Figure A6 for some examples from our trained model).

We note that there have been two prior works ([Jaiswal et al., 2018](#); [Upadhyay and Schrater, 2018](#)) that have combined capsules with GANs by changing the discriminator model to use (non-convolutional) capsules. They therefore were unable to demonstrate the benefits of using capsules in the generator model, and did not demonstrate an application to images beyond ( $64 \times 64$  pixels). In initial experiments (Appendix C), we found that standard convolutional discriminators ([Radford](#)

et al., 2015) qualitatively performed better than convolutional capsule discriminators, and so opted to use the former.

### 3. Methods

#### 3.1. Convolutional Capsules and Local Dynamic Routing

Our generator utilises convolutional capsules (LaLonde and Bagci, 2018), where the dynamic routing is instead preceded by a convolution instead of a weight matrix multiplication. Convolutional capsule layers take inputs  $a$ , of size  $[B, I, C_I, W_I, H_I]$ , and output  $\hat{u}$ , of size  $[B, I, W_J, H_J, J, C_J]$ , where  $B$  = batch size,  $I$  = number of input capsules,  $C_I$  = number of input channels,  $W$  = width,  $H$  = height,  $J$  = number of output capsules and  $C_J$  = number of output channels. In the convolution step, the kernels  $K_W \times K_H$  are shared between the input capsules in order to reduce the number of parameters, such that the total number of parameters per convolution is  $C_I \times K_W \times K_H \times C_J \times J$ . For each layer of the network, the output of the convolution  $\hat{u}$ —the activations of the child capsules—are routed to all the parent capsules. We use  $i$  and  $j$  to denote indices for child and parent capsules, respectively.

LaLonde and Bagci (2018) also introduce the local dynamic routing algorithm to accompany convolutional capsules. With convolutional capsules, since  $\hat{u}$  has width and height dimensions, each spatial location in the input capsule is routed to the same spatial location in the output capsules, using a spatial kernel of size  $K_r = 1$ <sup>1</sup>. The spatial kernel is formed from the parameter  $b$ , of size  $[I, W_J, H_J, J]$ , which is normalised to form vectors  $c_{ij}$  using the softmax function:

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_i \exp(b_{ij})}, \quad (4)$$

where  $b_{ij}$  is updated in every routing iteration.

The routing process involves taking the dot product of  $\hat{u}_{ij}$  with  $c_{ij}$  over the input capsules to give  $s_j$ —each spatial location ( $W, H$ ) in the input capsules is routed to the corresponding location in the output capsules.

The output of the dot product,  $s_j$ , is then passed through the nonlinear “squash” function to give  $v_j$ :

$$v_j = \frac{\|s_j\|_2^2}{1 + \|s_j\|_2^2} \cdot \frac{s_j}{\|s_j\|_2} \quad (5)$$

The purpose of this nonlinear function is to normalise the output between  $[0, 1]$ , and to ensure that the vector direction is retained, which is important in the  $b_{ij}$  update step. For the update,  $b_{ij}$  is incremented with the dot product of  $v_j$  with  $\hat{u}_{ij}$ , which is a key element of the dynamic routing algorithm. The dot product essentially looks at the similarity between the input and output capsules, and updates  $b_{ij}$  accordingly (similar features increase the value of  $b$ , and dissimilar features reduce it).

The entire procedure for convolutional capsules followed by dynamic routing is illustrated in Algorithm 1, where we use  $*$  to denote either convolution or transpose convolution.

---

1. While both LaLonde and Bagci (2018) and we use  $K_r = 1$ , it is possible to use any kernel size.

**Algorithm 1:** Convolutional Capsules + Local Dynamic Routing

---

**Input:**  $a$ , capsules in layer  $l$ ;  $l$ , layer;  $r$ , iterations; bias; weight  
**Output:**  $v_j$ , capsules in layer  $(l+1)$

$$\hat{u}_{I \times J \times C_J} \leftarrow \text{bias}_{J \times C_J} + \sum_{n=0}^{C_I} \text{weight}_{J \times C_J, n} * a_n$$

for all capsules  $i$  in layer  $l$  and capsules  $j$  in layer  $(l+1)$ :  $b_{ij} \leftarrow 0$

**for**  $l$  **to**  $r$  **do**

- for all capsules  $i$  in layer  $l$  and capsule  $j$  in layer  $(l+1)$ :  $c_{ij} \leftarrow \text{softmax}(b_{ij})$   $\triangleright$  Eq. 4
- for all capsules  $j$  in layer  $(l+1)$ :  $s_j \leftarrow \sum_i c_{ij} \hat{u}_{ij}$
- for all capsules  $j$  in layer  $(l+1)$ :  $v_j \leftarrow \text{squash}(s_j)$   $\triangleright$  Eq. 5
- for all capsules  $i$  in layer  $l$  and capsule  $j$  in layer  $(l+1)$ :  $b_{ij} \leftarrow b_{ij} + \hat{u}_{ij} \cdot v_j$

**end**

---

### 3.2. Convolutional Capsule Generator

Our generator architecture is based on the U-net-style encoder-decoder network with skip connections (Figure 1) (Ronneberger et al., 2015), similarly to pix2pix (Isola et al., 2017) and SegCaps (LaLonde and Bagci, 2018). This architecture is designed to output an image that is aligned with the structure of the input, and to encode both low and high level features by using skip connections. We later demonstrate that capsules learn highly relevant features at the last layer, while having far fewer parameters as compared to the original pix2pix.

In order to generate a distribution of synthetic images, we additionally augment our network with a 100D latent vector  $z \sim \mathcal{N}(0, \mathbb{I})$  (Figure 1). This vector is passed through a fully-connected layer of size  $100 \times 256^2$ , with the output reshaped to be the same spatial dimensions as the input image ( $256 \times 256$ ). This is then concatenated with the input segmentation label as an additional channel, so that the input into the first convolutional layer is of size  $[B, 2, 256, 256]$ . By fixing the input segmentation label and sampling different latent vectors, the network is able to produce varied images with the same geometry as represented by the label.

### 3.3. Conditional DCGAN Discriminator

Our discriminator is mainly based on the deep convolutional GAN (DCGAN) architecture (Radford et al., 2015). As per the original, the network is built with standard 2D convolutions followed by batch normalisation (Ioffe and Szegedy, 2015) and leaky ReLU nonlinearities (Xu et al., 2015); we also add a final fully-connected layer of size  $169 \times 1$  to compensate for the larger input size. To make  $D$  conditional, it receives as an input the image and label concatenated, giving an input size  $[B, 2, 256, 256]$ .

## 4. Datasets

We used both a real microscopy dataset and a physics-based dataset in our experiments (Figure 4). These datasets comprise of both labels and images. In addition, we also describe several methods that can take labels (from either dataset) as input, and generate new images.

#### 4.1. Synthetic Datasets

One baseline for synthesis of axon data is based on a simple statistical shape model. This model is produced by adding point locations – created by taking angle-constrained random walks – to a list of node locations in a 2D plane. The skeleton of the axon-like structures is obtained by fitting spline curves through these points (Wen and Chklovskii, 2008). This approach permits an infinite number of variations of axon-like geometry to be specified. Ground-truth binary labels for the corresponding centrelines are defined by keeping all pixels within a short distance from the curves, but the full curve description is maintained for a physics-based imaging model.

**PBAM-SSM:** Given a spline description, the corresponding images can be used to drive a simple physics-based imaging appearance model (PBAM). This model first builds a distance map between the spline description of the axon centreline, and the rest of the image. Distances are converted into intensity values by assuming a Gaussian axon profile or similar centreline-dependent, variable-width intensity profile. The final image is obtained by introducing intensity variations along the axons, optical blur, white and coloured noise sources. Synaptic protrusions are also added onto the axons. The combined geometry/appearance model is used as a baseline generative model for microscopy images.

**CapsPix2Pix-SSM & pix2pix-SSM:** The statistical shape model representing axons can be used to drive different geometrical realisations for the appearance-based generative models based on either the CapsPix2Pix or pix2pix networks.

#### 4.2. Microscopy Dataset

We combined data from two published sources (Bass et al., 2017; Carty et al., 2018) to obtain 152 ( $512 \times 512$ ) 2D images (produced from a max projection over 3D image stacks), and manually produced the corresponding labels; 20 of the images are held out for testing. These images were collected using in-vivo two-photon microscopy from the mouse somatosensory cortex. Examples of the labels and images in this dataset are shown in Figure 4. The labels are binary segmentation maps of the axons. The full dataset is available at <https://doi.org/10.5281/zenodo.2559237>.

#### 4.3. Training of Networks

The conditional generative models pix2pix and CapsPix2Pix are trained on labels and images from the microscopy dataset, which is augmented with random crops of size  $256 \times 256$ , which we denote as “cropped real” (CR). The segmentation models (U-nets) are trained using data synthesised from labels obtained from the SSM, and labels from the microscopy dataset, augmented with random flips, rotations, zooming and crops of size  $64 \times 64$ , which we denote as “augmented real” (AR). Examples of synthetic images from different models compared to real data are shown in Figure 5. We use 26,400 images in all our experiments for training the segmentation, and 13,200 images in all our experiments for training conditional generative models.

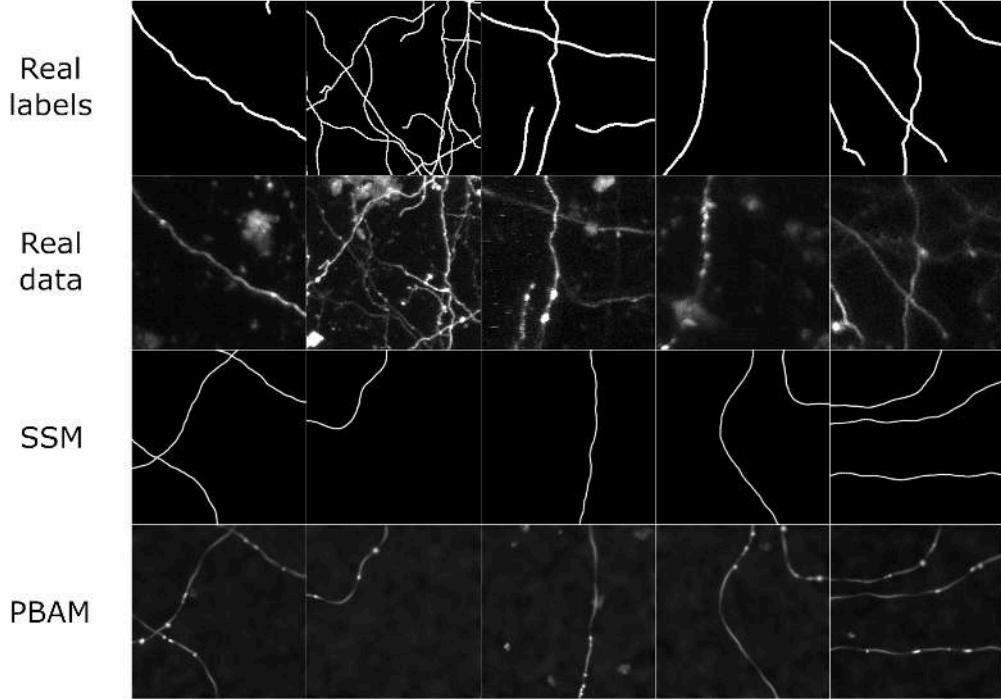


Figure 4: Examples of labels and images from the real and physics-based synthetic datasets. Abbreviations: SSM = statistical shape model; PBAM = physics-based imaging appearance model.

## 5. Experiments and Results

### 5.1. Evaluation of Generative Models

Evaluating the quality of generative models is known to be a difficult problem (Theis et al., 2016). Several quantitative methods, such as Parzen window log-likelihood estimates (Breuleux et al., 2010) and the Inception score (Salimans et al., 2016) have been proposed, but these may not always be appropriate. We provide a further discussion in Appendix B. In the case of conditional image synthesis, we could use quantitative metrics such as mean squared error or structural similarity index to compare against “ground truth” images. Unfortunately, these do not give a meaningful evaluation of the quality of the synthesis, since if the structure is similar, but the quality or data distribution is bad, the scores might still be high. A more general way of evaluating synthetic image quality is to allow human observers to discriminate between “real” and “fake” images. However, as our dataset contains biomedical images, subjects would need to be trained on the data first before performing the task. We instead choose to quantitatively evaluate our method by testing how its synthesised images affect the performance in downstream tasks of interest—in this case, segmentation. This task-based evaluation has previously been used for evaluating generative models in, for example,

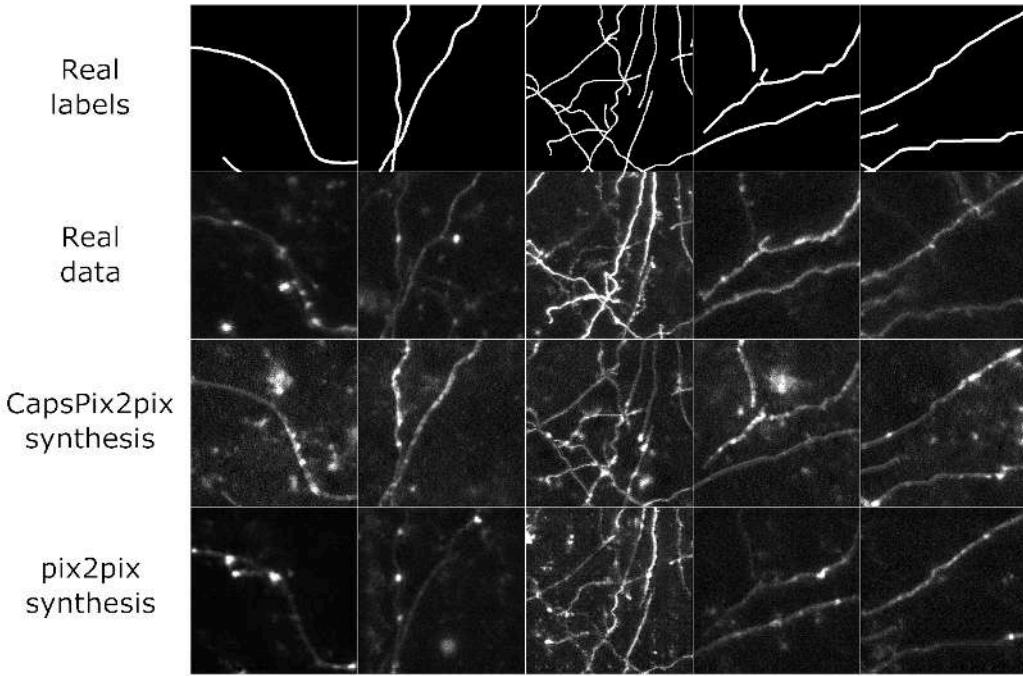


Figure 5: Examples of synthetic images compared to images from the real dataset. All images have a resolution of  $256 \times 256$  pixels.

skin lesion datasets ([Frid-Adar et al., 2018a,b](#)). Qualitatively, it is possible to examine the latent space to see how well the network has learned high level information about the data distribution, and the network’s activations to see what features have been learned.

## 5.2. Quantitative Analysis

As the end goal of our work on generative models is to improve the analysis of biomedical datasets, we chose to quantitatively evaluate our model by using synthesised images in a segmentation task. We trained separate U-nets from scratch, on different datasets (real, and synthetic images from different models), and repeat each experiment 10 times to get means and standard deviations per model. Testing was done on the same 400 crops ( $64 \times 64$ ) of held-out test dataset of 20 images ( $512 \times 512$ ). Performance was measured using the Dice score, receiver operator characteristic (ROC, Figure [A3](#)) area under the curve (AUC), and precision-recall (PR) AUC metrics (Table [1](#)). We compared our model, CapsPix2Pix, to the state-of-the-art pix2pix ([Isola et al., 2017](#)). ROC and test plots for selected U-nets (trained on different datasets) are displayed in Figures [A3](#) and [A4](#).

**Training U-net from scratch on real or synthetic datasets:** When comparing the Dice score per experiment (Table [A1](#)), we found that the CapsPix2Pix-SSM model performs better than the pix2pix-SSM model ( $p < 0.0001$ ,  $t$ -test), and better than the PBAM-SSM ( $p < 0.0001$ ,  $t$ -test).

Table 1: **Segmentation results.** The same U-net architecture was trained on different datasets, and evaluated on the test set of 400 crops of size  $64 \times 64$  from the original dataset (20 test images of size  $512 \times 512$ ). The same number of samples—26,400 images of size  $64 \times 64$ —were used in training in all experiments. Hyphenated names refer to image synthesis model, followed by the label source. The  $\dagger$  refers to the usage of only 1,320 unique labels, but generating 20 images per label. We highlight in bold the best scores in each section separated by a horizontal line. We repeat each experiment 10 times, and report averages and standard deviations across those experiments in this table. Abbreviations: PBAM = physics-based imaging appearance model; SSM = (synthetic labels of the) statistical shape model; AR = augmented real (labels).

Images	Labels	Pretrained	Dice	ROC AUC	PR AUC
PBAM	SSM	No	$0.6094 \pm 0.0083$	$0.9560 \pm 0.0054$	$0.6157 \pm 0.0102$
pix2pix	AR	No	$0.6592 \pm 0.0034$	$0.9670 \pm 0.0004$	$0.6834 \pm 0.0034$
pix2pix	SSM	No	$0.6353 \pm 0.0036$	$0.9631 \pm 0.0012$	$0.6635 \pm 0.0063$
CapsPix2Pix	AR	No	$0.6407 \pm 0.0040$	$0.9608 \pm 0.0022$	$0.6572 \pm 0.0050$
CapsPix2Pix	SSM	No	$0.6528 \pm 0.0028$	$0.9637 \pm 0.0014$	$0.6691 \pm 0.0002$
Real data	AR	No	<b><math>0.6827 \pm 0.0010</math></b>	<b><math>0.9725 \pm 0.0002</math></b>	<b><math>0.7149 \pm 0.0015</math></b>
pix2pix	AR $^\dagger$	No	$0.6438 \pm 0.0031$	<b><math>0.9652 \pm 0.0009</math></b>	$0.6676 \pm 0.0047$
pix2pix	SSM $^\dagger$	No	$0.6393 \pm 0.0041$	$0.9635 \pm 0.0014$	$0.6657 \pm 0.0045$
CapsPix2Pix	AR $^\dagger$	No	$0.6423 \pm 0.0075$	$0.9638 \pm 0.0015$	$0.6621 \pm 0.0117$
CapsPix2Pix	SSM $^\dagger$	No	<b><math>0.6540 \pm 0.0078</math></b>	$0.9620 \pm 0.0048$	<b><math>0.6689 \pm 0.0069</math></b>
Real data	AR	pix2pix-AR	$0.6856 \pm 0.0005$	$0.9731 \pm 0.0002$	$0.7170 \pm 0.0009$
Real data	AR	pix2pix-SSM	$0.6830 \pm 0.0005$	$0.9724 \pm 0.0002$	$0.7145 \pm 0.0007$
Real data	AR	CapsPix2Pix-AR	<b><math>0.6870 \pm 0.0005</math></b>	<b><math>0.9734 \pm 0.0001</math></b>	<b><math>0.7190 \pm 0.0006</math></b>
Real data	AR	CapsPix2Pix-SSM	$0.6855 \pm 0.0005$	$0.9730 \pm 0.0001$	$0.7175 \pm 0.0006$

However, the pix2pix-AR model performs better than CapsPix2Pix-AR ( $p < 0.0001$ ,  $t$ -test). We conclude that, overall, training U-net *from scratch* on CapsPix2Pix or pix2pix synthetic data leads to comparable results, since they each perform better in one case when using either SSM or AR as input labels. Overall, using the real data with augmentation leads to better segmentation results than training only on the synthetic images ( $p < 0.0001$ ,  $t$ -test).

**Pre-training U-net on synthetic datasets:** To test whether using the synthetic datasets could improve upon only training with real images, we pretrain our U-nets on synthetic images, and finetuned using the real data. Doing so significantly improves the Dice score (compared to just training on real data) when pretraining on pix2pix-AR data ( $p < 0.0001$ ,  $t$ -test), but not on pix2pix-SSM ( $p > 0.05$ ,  $t$ -test). In contrast, we achieve a significant improvement on both CapsPix2Pix-AR and CapsPix2Pix-SSM data (AR & SSM;  $p < 0.0001$ ,  $t$ -test). Overall, we reach the *best results* when pretraining on CapsPix2Pix-AR ( $0.6870 \pm 0.00005$ ), and we find that it is significantly better than the pretrained model on pix2pix-AR data ( $p < 0.0001$ ,  $t$ -test).

**Training U-net on synthetic datasets with reduced unique labels:** We performed an additional experiment to quantitatively demonstrate that our model can synthesise more diverse data, given the same input label. We trained U-nets with a reduced number of SSM or AR labels (1,320 unique labels), but several synthetic images per label (20 images) for a total of 26,400 (i.e. same number of images as in other experiments), for both CapsPix2Pix and pix2pix. We found that while training on CapsPix2Pix-SSM<sup>†</sup> data (where <sup>†</sup> refers to a model trained with reduced labels) leads to good performance, training on pix2pix-SSM<sup>†</sup> data leads to overfitting, and performance is reduced to  $\sim 0.61$  by the end of training (bottom of Figure A4). In addition, when comparing performance of the best models (i.e., before overfitting; Table 1), we found that CapsPix2Pix-SSM<sup>†</sup> is significantly better than pix2pix-SSM<sup>†</sup> and pix2pix-AR<sup>†</sup> ( $p < 0.0001$ , and  $p < 0.005$ ,  $t$ -test). This strongly indicates that our model is better able to synthesise diverse images. See Figure A5 for randomly picked examples of synthetic images from the same input labels.

### 5.3. Qualitative Analysis

We investigated how well our model captured the data distribution of the axon and noise using linear interpolation between 2 randomly sampled  $z$  vectors (Figure 3). The synthetic images display a large variation in noise and axon intensity (including synapses on different locations on the axon). We also display the features (activations) of the last capsule layer, demonstrating the variety of features learned by the network, even with a reduced number of parameters. In addition, we found that individual capsules appear to group similar features (Figure A6). This could be key in learning a balanced data distribution within and across classes, as different capsules could represent different classes, and the features within each capsule could encode the variability within that class.

We also compared the 16 features (the last layer activations) produced from the same label between pix2pix and CapsPix2Pix (Figure 2). While pix2pix also learns different features, many of them appear to be redundant and do not capture the noise very well (see Figure A7 for all pix2pix features at the last layer). CapsPix2Pix learns both features of the axon and noise (high intensity, and general noise) quite well.

Finally, we synthesised different images for the same label from each network (Figure A5), and found that while CapsPix2Pix can generate a large variation of images from one label, pix2pix is only able to make minor alterations to the images.

## 6. Conclusion

In this paper we introduced CapsPix2Pix, a novel method which utilises convolutional capsule networks in the cGAN framework for synthesising images conditional on segmentation labels.

We quantitatively validate our method by training U-nets on synthesised data from CapsPix2Pix and the state-of-the-art pix2pix model, which leads to comparable performance in a segmentation task (Section 5.2), while using far fewer parameters (Appendix A). More relevant to our end goal, we show that if the U-net is pretrained with synthesised data from CapsPix2Pix, it increases performance in the segmentation task. We attempt to compare the data distributions of the real and synthetic images via kernel density estimation, but conclude that this method is invalid in our case (see discussion in Appendix B).

Qualitatively, we show that CapsPix2Pix is able to synthesise considerably different images from the same images, while pix2pix does not (Section 5.3, Figure A5). We quantified whether being able to synthesise different images from the same label would impact the results when training

U-net. We synthesised images from a reduced number of unique labels with 20 images per label, and showed that segmentation performance was better when CapsPix2Pix data was used (Section 5.2, bottom of Figure A4). We also show that CapsPix2Pix learns relevant features for our dataset, and appears to capture the distribution of the noise and neuron classes well (Figures 2, 3 and A6), while pix2pix mainly just captures the axon class (Figures 2 and A7). Because of this property, we hypothesise that CapsPix2Pix could generalise well to other or more complicated datasets with multiple classes, and possibly even improve on the problem of mode collapse. For example, it could apply well to other biomedical datasets, where class imbalance is a common problem, and the dataset could benefit from a generator that captures a balanced distribution both within and across classes. In particular, our model’s ability to synthesise diverse images could be of greater importance when dealing with datasets with a small amount of labelled data—another common problem in biomedical datasets. For instance, Dai et al.( 2019) apply deep reinforcement learning to the problem of centreline tracking on axons (Bass et al., 2017), but are limited to training on hand-engineered synthetic data due to the amount of labelled data required; a natural next step to improve the results would be to use CapsPix2Pix to instead generate a large amount of more realistic data.

## Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/L016737/1].

## References

- Moustafa Alzantot, Supriyo Chakraborty, and Mani Srivastava. Sensegen: A deep learning architecture for synthetic sensor data generation. In *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*, pages 188–193. IEEE, 2017.
- Cher Bass, Pyry Helkkula, Vincenzo De Paola, Claudia Clopath, and Anil Anthony Bharath. Detection of axonal synapses in 3d two-photon images. *PloS one*, 12(9):e0183309, 2017.
- Christoph Baur, Shadi Albarqouni, and Nassir Navab. Melanogans: High resolution skin lesion synthesis with gans. *arXiv preprint arXiv:1804.04338*, 2018.
- Olivier Breuleux, Yoshua Bengio, and Pascal Vincent. Unlearning for better mixing. *Universite de Montreal/DIRO*, 2010.
- Alison J Canty, Johanna S Jackson, Lieven Huang, Antonio Trabalza, Cher Bass, Graham Little, and Vincenzo De Paola. Single-axon-resolution intravital imaging reveals a rapid onset form of wallerian degeneration in the adult neocortex. *bioRxiv*, page 391425, 2018.
- Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.
- Tianhong Dai, Magda Dubois, Kai Arulkumaran, Jonathan Campbell, Cher Bass, Benjamin Billot, Fatmatulzehra Uslu, Vincenzo de Paola, Claudia Clopath, and Anil Anthony Bharath. Deep

reinforcement learning for subpixel neural tracking. In *Medical Imaging with Deep Learning*, 2019.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *arXiv preprint arXiv:1803.01229*, 2018a.

Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 289–293. IEEE, 2018b.

Adrien Gaidon, Antonio Lopez, and Florent Perronnin. The reasonable effectiveness of synthetic visual data. *International Journal of Computer Vision*, 126(9):899–901, 2018.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

Stefan Hinterstoesser, Vincent Lepetit, Paul Wohlhart, and Kurt Konolige. On pre-trained image features and synthetic images for deep learning. *arXiv preprint arXiv:1710.10710*, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

Ayush Jaiswal, Wael AbdAlmageed, and Premkumar Natarajan. Capsulegan: Generative adversarial capsule network. *arXiv preprint arXiv:1802.06167*, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Dimitrios Korkinof, Tobias Rijken, Michael O’Neill, Joseph Yearsley, Hugh Harvey, and Ben Glocker. High-resolution mammogram synthesis using progressive generative adversarial networks. *arXiv preprint arXiv:1807.03401*, 2018.

Rodney LaLonde and Ulas Bagci. Capsules for object segmentation. *arXiv preprint arXiv:1804.04241*, 2018.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS Autodiff Workshop*, 2017.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1060–1069. JMLR.org, 2016a.

Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Advances in Neural Information Processing Systems*, pages 217–225, 2016b.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866, 2017.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

Leon Sixt, Benjamin Wild, and Tim Landgraf. Rendergan: Generating realistic labeled data. *Frontiers in Robotics and AI*, 5:66, 2018.

L Theis, A van den Oord, and M Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR 2016)*, pages 1–10, 2016.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

Yash Upadhyay and Paul Schrater. Generative adversarial network architectures for image synthesis using capsule networks. *arXiv preprint arXiv:1806.03796*, 2018.

Quan Wen and Dmitri B. Chklovskii. Costbenefit analysis of neuronal morphology. *Journal of Neurophysiology*, 2008.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

## Appendix A. Computational Comparison

Here we show a comparison of the computational footprint of CapsPix2Pix and pix2pix.

**Weights and activations:** We compared the number of trainable parameters for the generator of CapsPix2Pix and pix2pix. CapsPix2Pix has 7.9M parameters while pix2pix has in total 54M parameters, i.e. CapsPix2Pix is  $\times 7$  smaller. The size of the activations in the network is comparable between the two networks—with CapsPix2Pix having 18M elements for a single sample, and pix2pix having 16M elements.

**Run-time in training and inference:** We found that during training, running 1 training epoch for a batch size of 1, for both generator and discriminator, takes CapsPix2Pix 0.27 seconds and pix2pix 0.055 seconds, i.e. CapsPix2Pix is  $\sim \times 5$  slower than pix2pix. During inference, isolating 1 run through the generator for 1 image, we found that it takes CapsPix2Pix 0.058 seconds, and pix2pix 0.00445 seconds, i.e. CapsPix2Pix is  $\sim \times 13$  slower than pix2pix when comparing the generator only. These run-time experiments were performed on the same PC with a GeForce GTX 1080Ti GPU.

## Appendix B. Quantitative Metrics for Evaluating Generative Models

Ideally, we would be able to quantify how well samples from a generative model match samples from the real data distribution. This is one motivation for using Parzen window estimates (Breuleux et al., 2010)—in which a nonparametric kernel density estimator is fit to real data, giving the ability to evaluate log-likelihoods for model samples against the density estimator. We performed this evaluation, fitting 100 Gaussian kernels, with bandwidths ranging from 0.1 to 0.4, to 10000 samples of real data ( $32 \times 32$ ), and evaluated the log-likelihoods of 10 draws of 1000 samples of synthetic data (pix2pix, CapsPix2Pix). As a control, we also evaluated 10 draws of 1000 samples from left-out real data. The results can be seen in Figure A1: while CapsPix2Pix has a higher log-likelihood than pix2pix across all bandwidths, real data has the lowest log-likelihoods. This has been shown to occur before, and unfortunately invalidates its use as a quantitative metric (Theis et al., 2015).

Another metric—the Inception score—is based on the Kullback-Leibler divergence between the conditional label distribution,  $p(y|x)$ , and the marginal label distribution,  $p(y)$ , where the distributions are evaluated using a pretrained discriminative network (originally Inception-v3 trained on ImageNet data) (Salimans et al., 2016). A high score corresponds to generating meaningful objects (the entropy of  $p(y|x)$  is low) and a wide range of classes (the entropy of  $p(y)$  is high). While this metric may make sense for image datasets with a large number of classes with a balanced set of samples, it breaks down for a small number of classes with imbalanced data—as is the case for our axon dataset (Bass et al., 2017). In particular, generated samples with a low amount of noise could potentially score higher, although this is unrealistic.

The Fréchet inception distance compares the activations in a pretrained discriminative network (originally Inception-v3) of both real and generated samples (Heusel et al., 2017). This comparison is similar to Parzen window estimation, but occurs in a relevant feature space rather than on raw features. The feature space of the pretrained network is important. The image statistics of generic, real images are vastly different to those of medical images, requiring a network pretrained on medical imaging data. However, the previously described problem with noise re-occurs, as we typically train discriminative networks to be robust to noise in the input images.

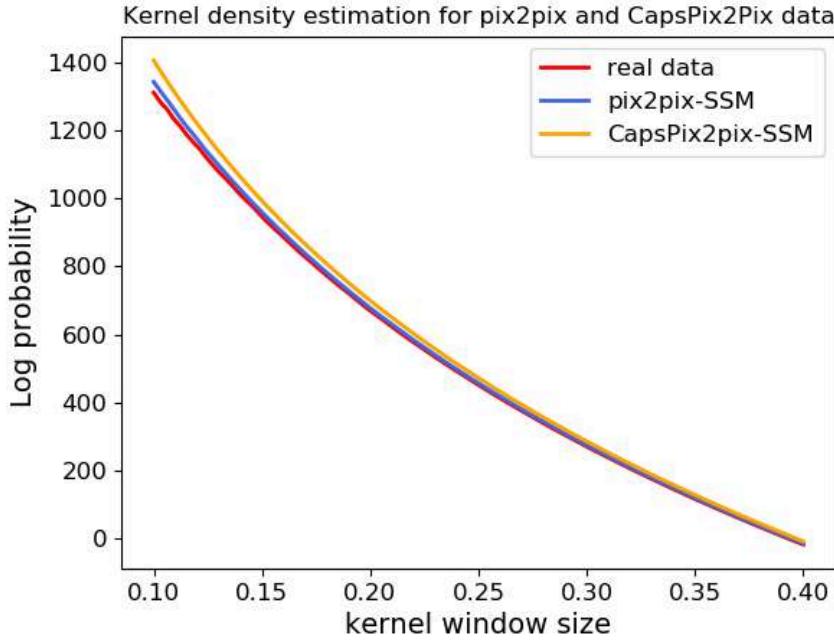


Figure A1: Kernel density estimation plot, comparing the average (of 10 experiments) log-likelihoods of samples from pix2pix, CapsPix2Pix and held-out real data. As the estimated log-likelihood of the real data is consistently below generated data, this invalidates its use as a quantitative metric. Abbreviation: SSM = (synthetic labels of the) statistical shape model.

### Appendix C. Additional Training Experiments For CapsPix2Pix

We performed some initial experiments to explore possible configurations for CapsPix2Pix. We tested whether using the same L1 lambda ( $\lambda = 0.1$ ) as in pix2pix (Isola et al., 2017) would work well for CapsPix2Pix. We found that while the synthetic images were reasonable, the images were too noisy at times, and might be unrealistic in some cases (Figure A2). Also, we found that during training, the quality of synthetic images oscillated a lot more than when using  $\lambda = 1$ . We also experimented with using different discriminator networks, including a network based on the traditional capsules (Sabour et al., 2017; Jaiswal et al., 2018; Upadhyay and Schrater, 2018), and one based on convolutional capsules (this network was similar in structure to a DCGAN, i.e., using batch normalisation, leaky ReLUs, the same number of convolutional layers, etc). We found that the traditional capsule discriminator did not work well at training the generator network, and that the convolutional capsule discriminator led to reasonable synthetic image generation, but was not as good as when using a DCGAN discriminator (Figure A2). In addition, both capsule-based discriminator networks increased training time. Lastly, we experimented with different ways to insert noise into the network, including broadcasting noise, adding it to the bottleneck, or using dropout in inference as in pix2pix, but these initial experiments were not promising.

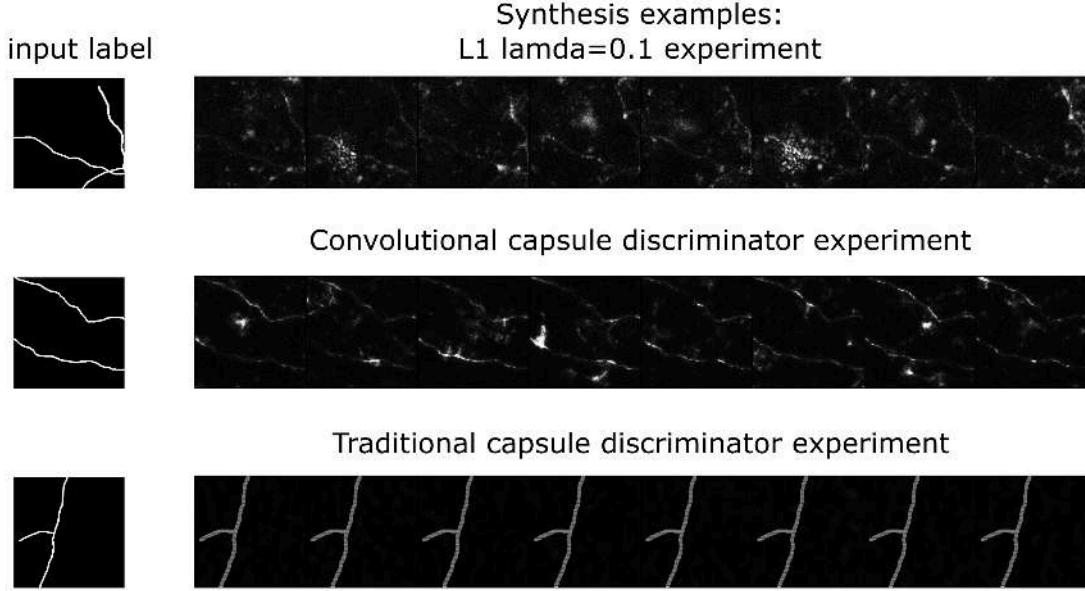


Figure A2: Additional training experiments.

## Appendix D. CapsPix2Pix Training Details

To optimise our network, we follow the approach laid out by [Goodfellow et al. \(2014\)](#). We train  $G$  to maximize  $\log D(x, G(x, z))$ , instead of minimizing  $\log(1 - D(x, G(x, z)))$ . In addition, we use Adam ([Kingma and Ba, 2014](#)) as our optimiser, with learning rate = 0.0002,  $\beta_1 = 0.5, \beta_2 = 0.999$ ; we linearly decay the learning rate to 0 by the end of training. To prevent  $D$  from becoming over-confident we use two-sided label smoothing ([Salimans et al., 2016](#)), with positive labels set to 0.9 and negative labels set to 0.1 when updating  $D$ . We also train  $G$  with a dropout probability of 0.5, which we do not apply during inference. Due to memory constraints we use small minibatch sizes (1-4) and hence do not use batch normalisation to train  $G$ .<sup>2</sup> We train  $D$  with batch normalisation, as in the original DCGAN. Unlike pix2pix, we input a latent vector  $z$ , so we do not need to add additional stochasticity to our generator during inference.

We trained the networks on the CR dataset (Section 4). Examples of the synthetic images are shown in Figure 5. The same dataset was used to train pix2pix for comparison. We trained pix2pix as described in the original work ([Isola et al., 2017](#)). All experiments were implemented using PyTorch ([Paszke et al., 2017](#)).

## Appendix E. U-net Segmentation Training Details

For each experiment we trained a standard U-net (with same padding so that the output has the same image size). These were trained on 26,400 ( $64 \times 64$ ) crops from various datasets, and were tested on the same 400 crops from the test dataset (20 images,  $512 \times 512$ ). During training we used a dropout

2. We found that using batch normalisation is effective when training on smaller images [ $64 \times 64$ ], where using higher batch size is possible.

probability of 0.5, a batch size of 32, and Adam with learning rate = 0.00001,  $\beta_1 = 0.5$ ,  $\beta_2 = 0.999$ . We keep aside 20% of the training data for validation.

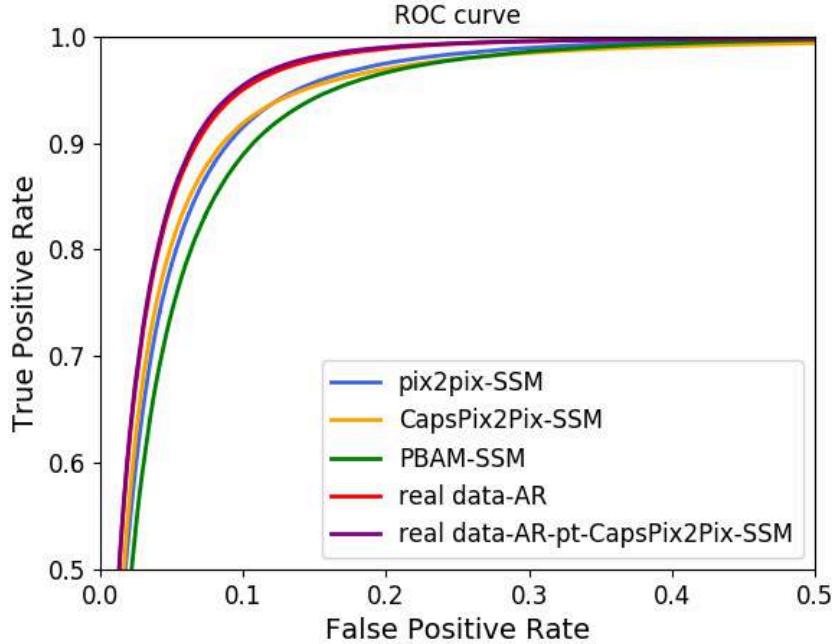


Figure A3: Comparison of U-net test results using ROC curves for different datasets. Each ROC curve is a concatenation of all false positive rates and true positive rates per dataset ( $\times 10$  experiments). Abbreviations: PBAM = physics-based imaging appearance model; SSM = (synthetic labels of the) statistical shape model; AR = augmented real (labels); pt = pretrained network.

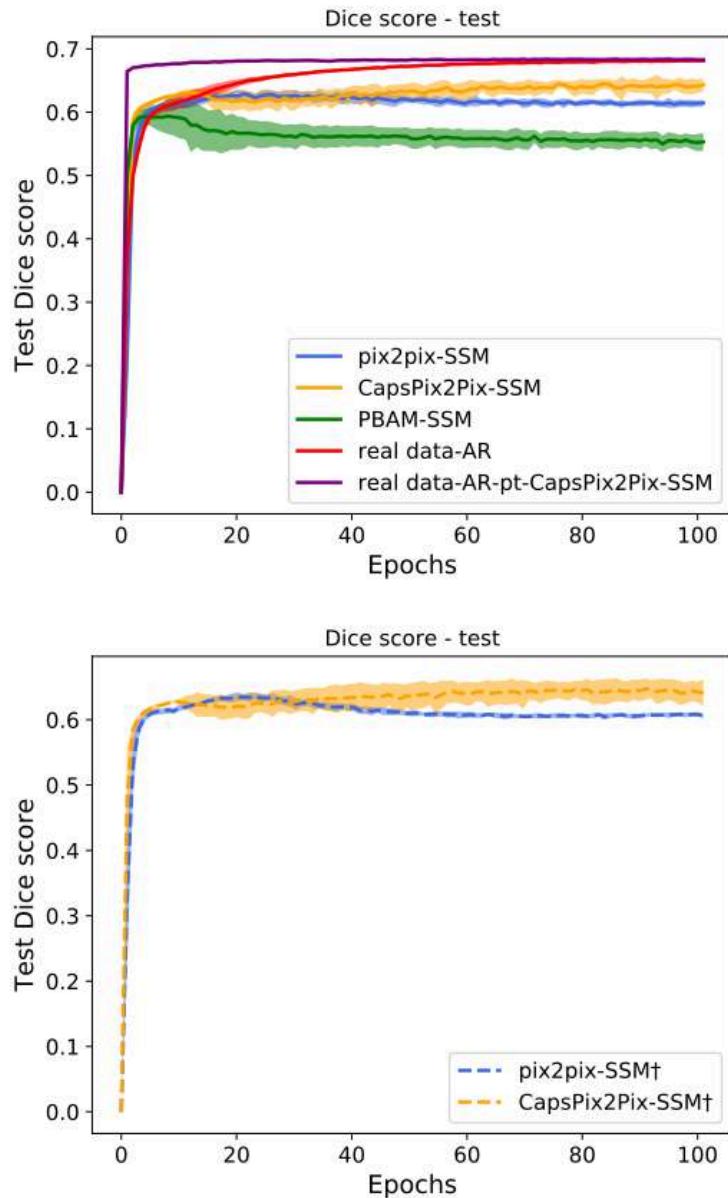


Figure A4: Comparison of U-net test curves for different datasets. The shaded regions represent  $\pm 1$  standard deviation. Abbreviations: PBAM = physics-based imaging appearance model; SSM = (synthetic labels of the) statistical shape model; AR = augmented real (labels); pt = pretrained network.

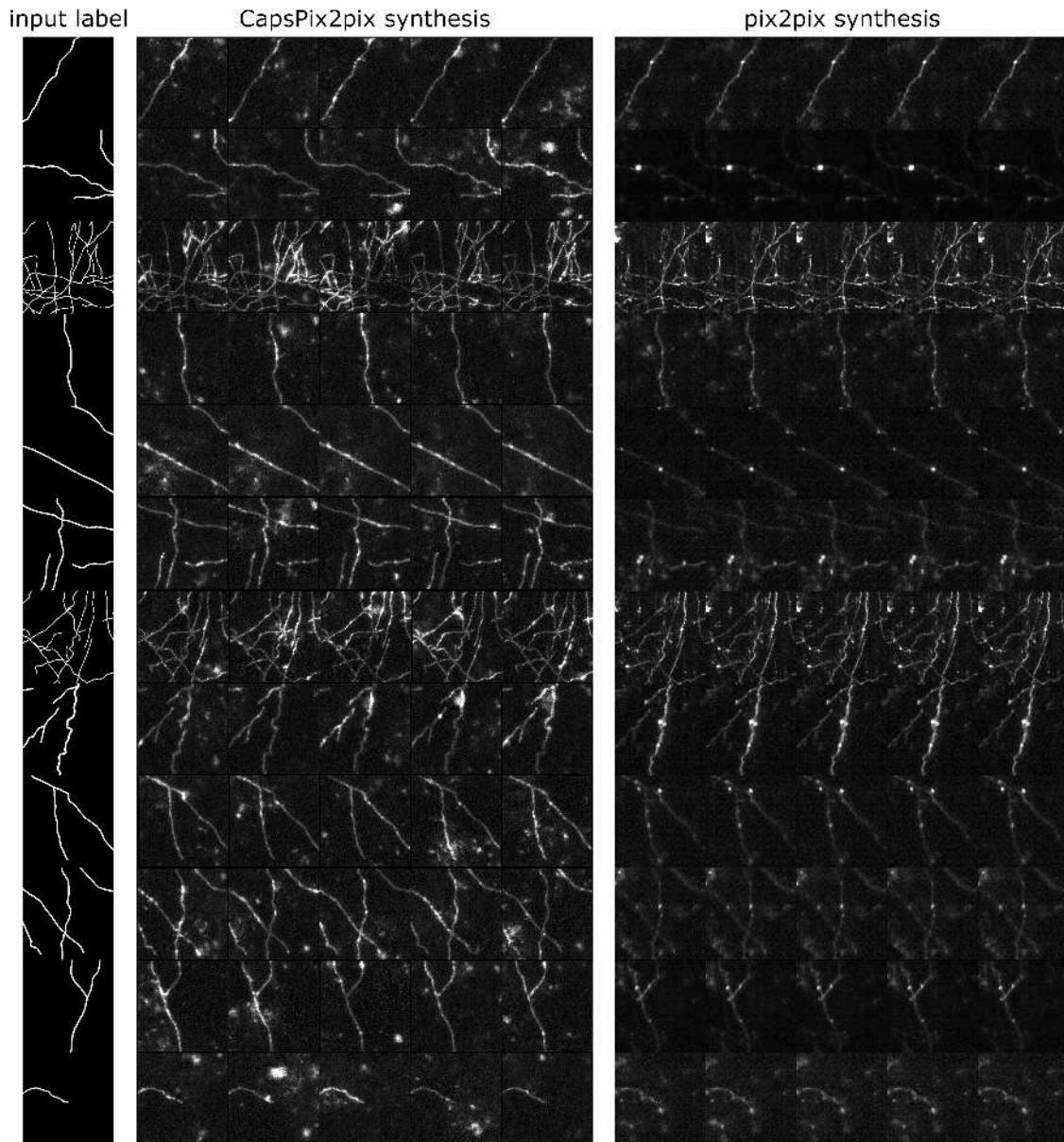


Figure A5: Examples of synthetic images from CapsPix2Pix and pix2pix based on the same labels. CapsPix2Pix has a large variation between synthetic images (when varying the  $z$  vector), but pix2pix only slightly alters the image (by applying dropout during inference).

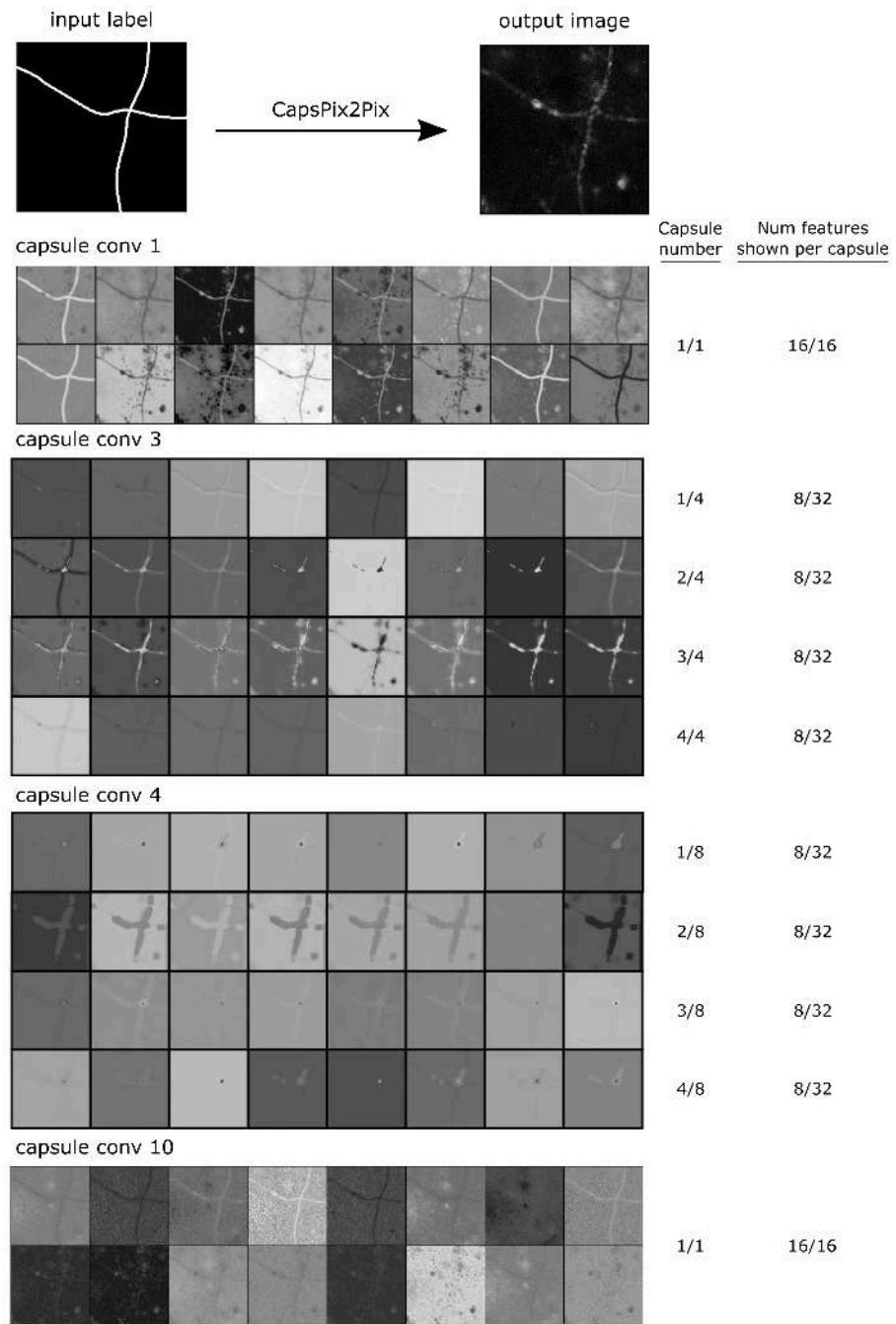


Figure A6: Examples of features (activations) at different capsule layers. The capsules from the intermediate layers group similar features (capsule conv 3 & 4). The outputs of the last convolutional capsules (capsule conv 10) are a combination of different types of features.

Table A1: Per-experiment Dice scores for different datasets. Abbreviations: AR = augmented real data; P2P = pix2pix; Caps = CapsPix2Pix; pt = pretrained on CapsPix2Pix-AR data; Std = standard deviation.

Exp	AR	PBAM-SSM	P2P-AR	Caps-AR	P2P-SSM	Caps-SSM	AR-pt
1	0.6834	0.6121	0.6564	0.6358	0.6448	0.6558	0.6869
2	0.6829	0.5954	0.6578	0.6350	0.6332	0.6531	0.6867
3	0.6828	0.6154	0.6606	0.6398	0.6418	0.6537	0.6871
4	0.6839	0.6071	0.6602	0.6400	0.6423	0.6500	0.6864
5	0.6818	0.6021	0.6564	0.6389	0.6475	0.6529	0.6882
6	0.6821	0.5993	0.6583	0.6343	0.6394	0.6503	0.6876
7	0.6827	0.6089	0.6617	0.6270	0.6411	0.6490	0.6864
8	0.6808	0.6169	0.6568	0.6329	0.6394	0.6556	0.6867
9	0.6843	0.6129	0.6678	0.6342	0.6353	0.6578	0.6866
10	0.6820	0.6240	0.6560	0.6350	0.6426	0.6500	0.6875
Mean	0.6827	0.6094	0.6592	0.6353	0.6407	0.6528	0.6870
Std	0.0010	0.0083	0.0034	0.0036	0.0040	0.0028	0.0005

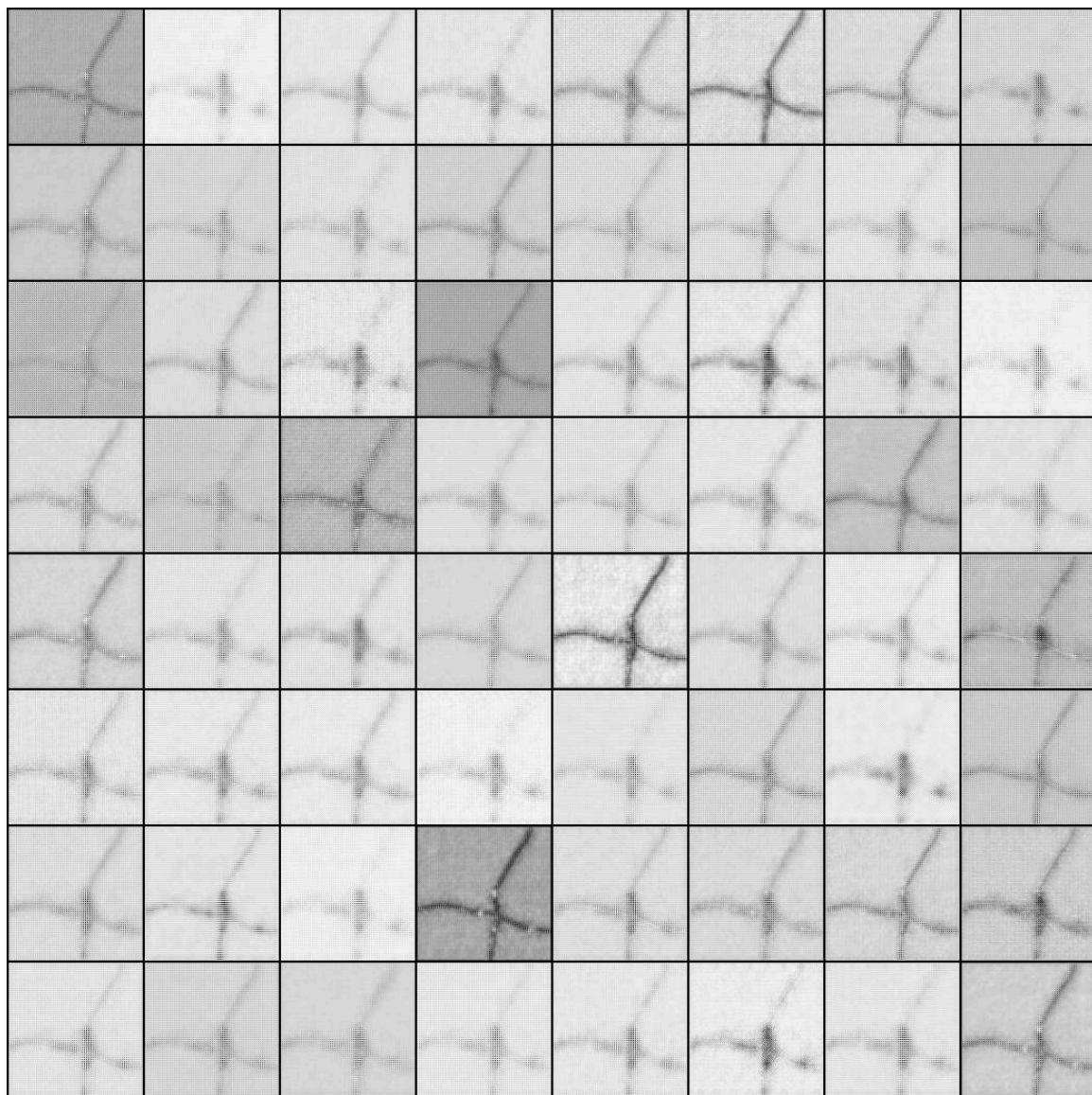


Figure A7: All 64 features (activations) of the last layer in a trained pix2pix model for a single input (the same as in Figure 2).

# Fusing Unsupervised and Supervised Deep Learning for White Matter Lesion Segmentation

**Christoph Baur<sup>1</sup>**

C.BAUR@TUM.DE

**Benedikt Wiestler<sup>3</sup>**

**Shadi Albarqouni<sup>1</sup>**

**Nassir Navab<sup>1,2</sup>**

<sup>1</sup> Computer Aided Medical Procedures (CAMP), TU Munich, Germany

<sup>2</sup> Whiting School of Engineering, Johns Hopkins University, Baltimore, United States

<sup>3</sup> Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, TU Munich, Germany

## Abstract

Unsupervised Deep Learning for Medical Image Analysis is increasingly gaining attention, since it relieves from the need for annotating training data. Recently, deep generative models and representation learning have lead to new, exciting ways for unsupervised detection and delineation of biomarkers in medical images, such as lesions in brain MR. Yet, Supervised Deep Learning methods usually still perform better in these tasks, due to an optimization for explicit objectives. We aim to combine the advantages of both worlds into a novel framework for learning from both labeled & unlabeled data, and validate our method on the challenging task of White Matter lesion segmentation in brain MR images. The proposed framework relies on modeling normality with deep representation learning for Unsupervised Anomaly Detection, which in turn provides optimization targets for training a supervised segmentation model from unlabeled data. In our experiments we successfully use the method in a Semi-supervised setting for tackling domain shift, a well known problem in MR image analysis, showing dramatically improved generalization. Additionally, our experiments reveal that in a completely Unsupervised setting, the proposed pipeline even outperforms the Deep Learning driven anomaly detection that provides the optimization targets.

**Keywords:** Deep Learning, Anomaly Detection, Unsupervised, Semi-Supervised, Supervised, White Matter Lesion Segmentation, Multiple Sclerosis

## 1. Introduction

Deep Learning for medical image analysis is still impeded by a general lack of labeled training data. Especially for medical image segmentation, the creation of pixel-level annotations is a very tedious, time-consuming and costly task, which often has to be carried out by domain experts. Although it has been shown that in some cases supervised models can be trained from very small training datasets (Ronneberger et al., 2015), usually large amounts of labeled training data are required to achieve compelling model performance. This is also the case for automatic segmentation of white matter lesions (WML) in brain MR images. WML, a result of demyelination of cells in the white matter of the brain, are important biomarkers for underlying degenerative neurological diseases such as Multiple Sclerosis and can vary greatly in size, shape and location (Carass et al., 2017). Supervised Deep Learning based WML segmentation methods (Brosch et al., 2016; Valverde et al., 2017; Roy et al., 2018) do not only have to cope with this wide variety of lesion appearances, but

additionally are confronted with the problem of domain shift: In contrast to CT data, intensities in MR images do not have a clear physical interpretation, and generally there is a discrepancy between data distributions of images produced with different MR scanners. It is this discrepancy which makes segmentation methods in MR data hardly generalize to new devices, and labeled training data from different scanners to deal with this issue might not be readily available.

To generally overcome these burdens, the community has made numerous efforts towards Unsupervised and Semi-Supervised Deep Learning, i.e. learning without any labeled data and learning from both labeled and unlabeled data, respectively. A promising approach towards this direction is pseudo-labelling (Lee, 2013), where supervised models are fine-tuned from labeled data together with unlabeled samples, for which labels have been predicted with the same model. Another approach, however limited to patch-based classification, are so-called Ladder networks (Rasmus et al., 2015). More recent works leveraged adversarial networks for Domain Adaptation by either explicitly enforcing domain invariant feature representations (Kamnitsas et al., 2017) or encouraging the model to also produce realistic segmentation masks on unlabeled samples (Dong et al., 2018). (Ganaye et al., 2018) employ semantic constraints to improve robustness of a brain structure segmentation model and (Jiang et al., 2018) use tumor-aware MR image synthesis from CT images to train a model for tumor segmentation from both labeled and synthetic data. At the example of MS lesion segmentation, (Baur et al., 2017) proposed a Semi-Supervised Deep Learning framework for Domain Adaptation of fully convolutional segmentation networks by encouraging domain invariant feature representations on randomly sampled embeddings.

A recent trend to overcome the burden of pixel-level annotations is to leverage deep generative models and deep representation learning for the task of Unsupervised Anomaly Detection (UAD) in medical images. Under the assumption that “healthy” data is readily available at hospitals, these approaches model the distribution of healthy anatomy and try to detect anomalies as outliers from the modeled distribution directly in image space. In early work (Schlegl et al., 2017), GANs were proposed to detect anomalies in small retinal OCT patches. For head CT, Sato et al. (2018) showed promising initial results using 3D Auto-Encoders and Pawlowski et al. (2018) studied the effects of averaging multiple Monte-Carlo dropout reconstructions in Bayesian Auto-Encoders for anomaly detection. For UAD in brain MR images, Chen and Konukoglu (2018) showed promising results for detecting large lesions with Constrained Adversarial AEs, and (Baur et al., 2018) found that spatial Auto-Encoders enable UAD at high resolution, ultimately allowing such models to also detect small MS lesions.

We propose a novel framework for WML segmentation that can benefit from both labeled and unlabeled data. Therefore, we combine i) a spatial Auto-Encoder, which performs UAD in brain MR images, and ii) a supervised segmentation network. We show that, in addition to labeled data, the anomaly detections obtained from the Auto-Encoder on unlabeled data can be leveraged for Unsupervised Domain Adaptation. As a proof-of-concept, we also show that the segmentation network can be trained from UAD results alone, which performs considerably better than the actual UAD approach, leaving us with a novel approach for Unsupervised Deep Learning as well.

## 2. Methodology

### 2.1. Overall Concept

Our framework consists of an Auto-Encoder (AE), which is used for UAD, and a UNet-like model for supervised image segmentation (Figure 1). In a first step, the AE is optimized for compressing

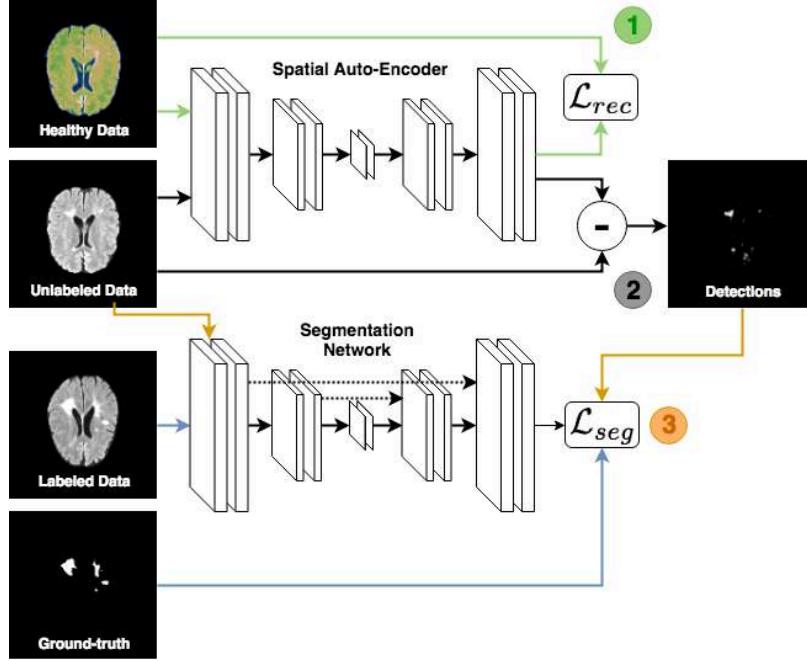


Figure 1: The proposed framework at a glance. Step 1: Training of a spatial AE on healthy data; Step 2: Inference on unlabeled data to obtain delineations; Step 3: Training of a supervised model from both labeled data with ground-truth and unlabeled data with UAD delineations.

and reconstructing images of healthy anatomy. Afterwards (step 2), it is used to detect and delineate anomalies in previously unseen, unlabeled data. In step 3, the UNet is trained in a supervised manner for pixel-wise WML, by jointly using labeled training data  $\mathcal{X}_L$  with ground-truth  $\mathcal{Y}_L$  as well as the unlabeled training data  $\mathcal{X}_U$ , for which the anomaly detection provides an “artificial ground-truth”  $\mathcal{S}$ .

## 2.2. Capturing normality for anomaly detection

Similar to (Baur et al., 2018), we train a 2D spatial Auto-Encoder to capture the notion of anatomical brain normality (see Figure 2 for a depiction of the network architecture). Given a set of healthy training data  $\mathcal{X}_H$ , we therefore optimized an AE for the following reconstruction objective:

$$\mathcal{L}_{rec}(\mathbf{x}, \hat{\mathbf{x}}) = \ell_1(\mathbf{x}, \hat{\mathbf{x}}) + \ell_2(\mathbf{x}, \hat{\mathbf{x}}) + \lambda_{gdl} gdl(\mathbf{x}, \hat{\mathbf{x}}) \quad (1)$$

The terms  $\ell_1$  and  $\ell_2$  constitute the pixel-wise Manhattan and Euclidean distances between input image  $\mathbf{x} \in \mathcal{X}_H$  and reconstruction  $\hat{\mathbf{x}}$ . In contrast to (Baur et al., 2018), which used an adversarial network to promote the reconstruction of realistic images, we used

$$gdl(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{i,j} |||x_{i,j} - x_{i-1,j}| - |\hat{x}_{i,j} - \hat{x}_{i-1,j}||| + |||x_{i,j-1} - x_{i,j}| - |\hat{x}_{i,j-1} - \hat{x}_{i,j}||| \quad (2)$$

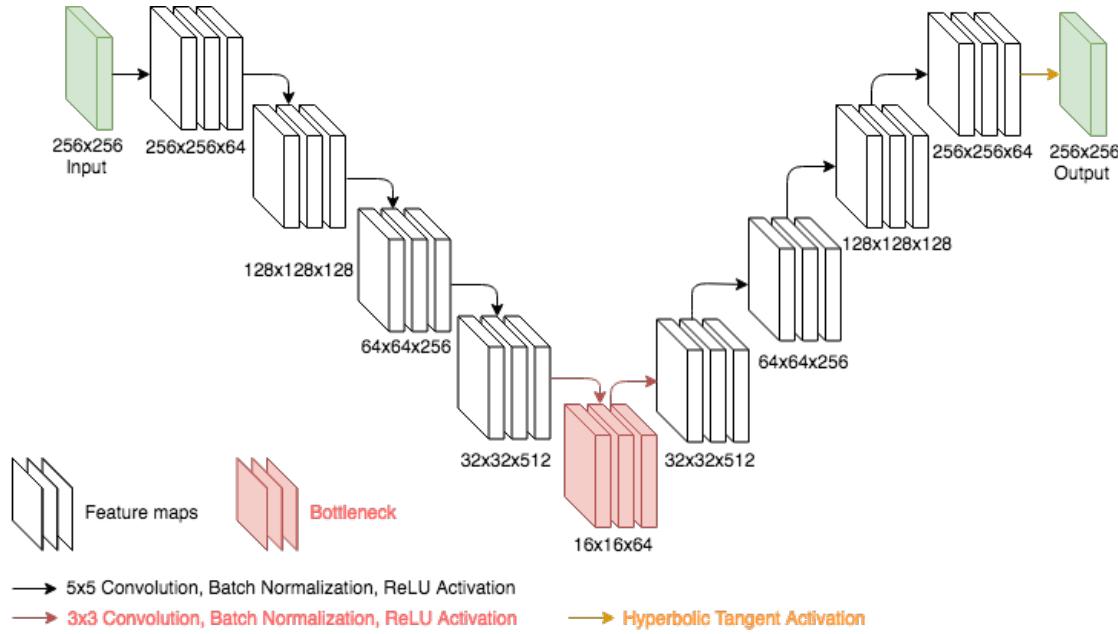


Figure 2: The architecture of the 2D Auto-Encoder used in our experiments.

the so-called gradient difference loss (Mathieu et al., 2015), weighted by  $\lambda_{gdl}$ . All these losses combined encourage the model not only to reconstruct coherent, but also very crisp images.

For detecting anomalies in a query sample  $\mathbf{x}^* \notin \mathcal{X}_H$  not being part of the Auto-Encoders training set,  $\mathbf{x}^*$  is propagated through the model and a pixel-wise residual between  $\mathbf{x}^*$  and its reconstruction  $\hat{\mathbf{x}}^*$  is computed:

$$\mathbf{r} = m(\max(\mathbf{x}^* - \hat{\mathbf{x}}^*, 0))$$

where  $m(\cdot, \cdot)$  is a non-linear  $5 \times 5 \times 5$  median filter for emphasizing connected anomalous structures and simultaneously removing unwanted, small residual pixels which might be high in intensity and lead to increased False Positive (FP) responses. Importantly, many of such potential FP residuals are already avoided by optimizing for the gdl-term, but the filtering is necessary. Further, we set any negative residuals to zero using ReLU to avoid detections of anomalies which do not resemble white matter lesions, as such lesions are usually hyper-intense in the FLAIR images we use. The resulting residuals are further binarized into images  $\mathbf{s}$  via thresholding, i.e.

$$\mathbf{s} = \mathbf{r} \geq t$$

and collected in a set of anomaly labels  $\mathcal{S}$ . How we choose the threshold  $t$  is explained in the experiments section.

### 3. Experiments and Results

#### 3.1. Dataset

For our experiments, we make use of two different datasets. We utilize the labeled data provided in the publicly available MICCAI 2008 MS lesion segmentation challenge dataset. The data ac-

quired at University of North Carolina ( $\mathcal{D}_{UNC}$ ) and the Children’s Hospital Boston ( $\mathcal{D}_{CHB}$ ) comprise FLAIR, T1 and T2-weighted images from 10 subjects per site.

Further, we use a non-public dataset, generously provided by our clinical partners at Klinikum Rechts der Isar, consisting of FLAIR and T1-weighted MR acquisitions of 68 healthy subjects ( $\mathcal{D}_{healthy}$ ) as well as 49 subjects which were diagnosed with MS ( $\mathcal{D}_{MS}$ ). For the latter, expert delineations of MS lesions were provided. All images have been acquired with a Philips Achieva 3T scanner.

**Preprocessing** Prior to any Deep Learning, all acquisitions have been projected to the SRI24 ATLAS space (Rohlfing et al., 2009), denoised using CurvatureFlow, Skull-Stripped using ROBEX (Iglesias et al., 2011) and normalized into the range of [0; 1]. While we train our models only from FLAIR images, we utilize the T1-modalities for co-registration and skull-stripping. Table 1 provides details about our training, validation and testing splits on the respective datasets.

Table 1: Training, Validation & Testing subjects of our datasets as well as additional subjects which are considered unlabeled in our experiments

<b>Dataset</b>	<b>Train</b> $\mathcal{X}_L$	<b>Val</b> $\mathcal{X}_{VAL}$	<b>Test</b> $\mathcal{X}_{TEST}$	<b>Additional</b> $\mathcal{X}_U$
$\mathcal{D}_{healthy}$	68	-	-	-
$\mathcal{D}_{MS}$	15	5	10	19
$\mathcal{D}_{CHB}$	6	2	2	-
$\mathcal{D}_{UNC}$	6	2	2	-

We utilize all FLAIR images of the healthy subjects in  $\mathcal{D}_{healthy}$  for training the AE. Further, we randomly split  $D_{CHB}$  and  $D_{UNC}$  each into training, validation and testing subjects. Similarly, out of  $\mathcal{D}_{MS}$  we utilize 15 randomly chosen subjects and their ground-truth segmentations for training a supervised segmentation model, 5 for validation and 10 subjects to test the models performance. We consider the remaining 19 subjects as unlabeled and utilize our AE to obtain an artificial “ground-truth”  $\mathcal{S}$ . The 5 labeled validation subjects with MS lesions are also used to choose an operating point for the UAD.

### 3.2. Auto-Encoder

A spatial AE, referred to as *UAD*, has been trained for 150 epochs from entire axial MR slices ( $256 \times 256$ px)  $\in \mathcal{D}_{healthy}$  with a learning-rate of 0.01. For the first 30 epochs, we set  $\lambda_{gdl} = 0.0$  to allow the model to converge to coherent reconstructions, and then set it to 100.0 to make the model focus more on reconstructing fine details. Afterwards, the MS lesion validation set  $\mathcal{X}_{VAL} \in \mathcal{D}_{MS}$  is processed by the AE to determine an Operating Point (OP)  $t = 0.0187$  which maximizes the DICE-Score on  $\mathcal{X}_{VAL}$ . In succession, all non-empty slices (determined via skull-stripping) of the additional 19 “unlabeled” subjects have been processed with the UAD model to detect anomalies, i.e. to generate our artificial ground-truth  $\mathcal{S}$  (see Figure 3 for an anomaly detection example from this set).

Similarly, another model  $UAD_{no-gdl}$  has been trained with fixed  $\lambda_{gdl} = 0.0$  throughout the entire 150 training epochs to study the impact of the gradient-difference-loss component. The anomaly detection performance of the two models on the testing subjects  $\mathcal{X}_{TEST} \in \mathcal{D}_{MS}$  is reported in Table 2.

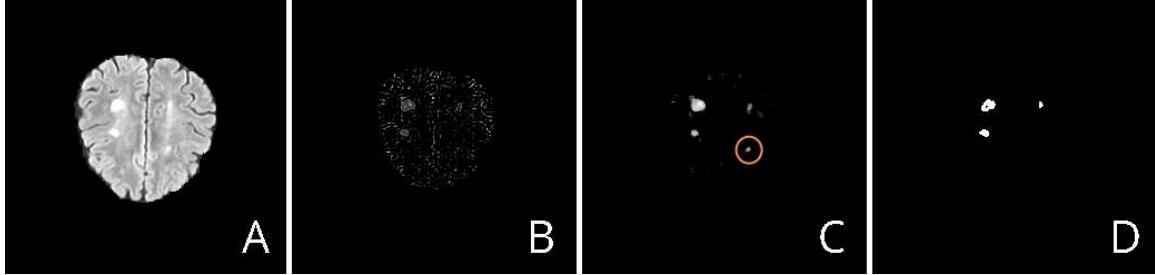


Figure 3: Anomaly detections provided by the AE. A: Input slice; B: Unprocessed Residuals; C: Postprocessed Residuals (with a FP encircled in orange); D: Ground-truth segmentation(s)

Notably, the UAD model performs better than its ablated counterpart  $\text{UAD}_{no-gdl}$  in all measures, which shows that the training with the gradient-difference-loss is indeed beneficial for anomaly detection.

### 3.3. Unsupervised and Semi-Supervised Deep Learning

To investigate the general suitability of our framework for Unsupervised and Semi-Supervised DL, we first conducted a set of experiments on data from a single domain. Therefore, we trained multiple UNet segmentation models using the  $\mathcal{D}_{MS}$  dataset, for which a larger number of subjects was available. The models comprise i) a supervised model  $\mathbf{A}_{\mathcal{Y}_L}$ , trained only on labeled data  $(\mathcal{X}_L, \mathcal{Y}_L) \in \mathcal{D}_{MS}$ , ii) a supervised model  $\mathbf{A}_{\mathcal{Y}_L+\mathcal{Y}_U}$ , trained with  $(\mathcal{X}_L, \mathcal{Y}_L)$  as well as the additional data  $(\mathcal{X}_U, \mathcal{Y}_U) \in \mathcal{D}_{MS}$  with its real ground-truth  $\mathcal{Y}_U$ , and iii) an unsupervised model  $\mathbf{A}_{\mathcal{S}}$ , trained only from additional “unlabeled” data  $\mathcal{X}_U$  and artificial ground-truth  $\mathcal{S}$ . Further, we trained a semi-supervised model  $\mathbf{A}_{\mathcal{Y}_L+\mathcal{S}}$  with  $(\mathcal{X}_L, \mathcal{Y}_L)$  and  $(\mathcal{X}_U, \mathcal{S})$ . All models have been trained for 50 epochs from  $128 \times 128$ px sized patches extracted around MS lesions. The OP is again determined on the validation data. Performances of the models on the testing set are reported in Table 2.

Table 2: Unsupervised, Semi-Supervised and Supervised Deep Learning experiments. Note: DICE is the overall Dice-Score, DICE ( $\mu \pm \sigma$ ) is the statistics over the Dice-Scores obtained per subject and AUPRC is the Area under the Precision-Recall-Curve

Model	DICE	DICE ( $\mu \pm \sigma$ )	AUPRC	Training Subjects
UAD	0.6343	$0.6156 \pm 0.0972$	0.6157	all in $\mathcal{D}_{healthy}$
$\text{UAD}_{no-gdl}$	0.6101	$0.5831 \pm 0.0989$	0.5989	all in $\mathcal{D}_{healthy}$
$\mathbf{A}_{\mathcal{Y}_L}$	0.7259	$0.7026 \pm 0.0635$	0.7537	15 $(\mathcal{X}_L, \mathcal{Y}_L)$
$\mathbf{A}_{\mathcal{S}}$	0.6792	$0.6643 \pm 0.0775$	0.6964	19 $(\mathcal{X}_U, \mathcal{S})$
$\mathbf{A}_{\mathcal{Y}_L+\mathcal{S}}$	0.7057	$0.6815 \pm 0.0743$	0.7254	15 $(\mathcal{X}_L, \mathcal{Y}_L)$ + 19 $(\mathcal{X}_U, \mathcal{S})$
$\mathbf{A}_{\mathcal{Y}_L+\mathcal{Y}_U}$	<b>0.7338</b>	<b><math>0.7148 \pm 0.0591</math></b>	<b>0.7642</b>	15 $(\mathcal{X}_L, \mathcal{Y}_L)$ + 19 $(\mathcal{X}_U, \mathcal{Y}_U)$

### 3.4. Semi-Supervised Domain Adaptation

Next, we investigated the suitability of our approach for the task of Domain Adaptation by comparing a semi-supervised model against supervised baselines. Therefore, we trained a supervised model  $\mathbf{B}_{\mathcal{Y}_L}$  using the training set  $\mathcal{X}_L \in \mathcal{D}_{CHB}$ . We further trained a semi-supervised model  $\mathbf{B}_{\mathcal{Y}_L+\mathcal{S}}$  using  $\mathcal{X}_L$  as well as the unlabeled data  $\mathcal{X}_U \in \mathcal{D}_{MS}$  with artificial labels  $\mathcal{S}$ , and a supervised upper bound model  $\mathbf{B}_{\mathcal{Y}_L+\mathcal{Y}_U}$  using the same data  $\mathcal{X}_L$  and  $\mathcal{X}_U$ , however optimizing for the real ground-truth  $\mathcal{Y}_U$  of  $\mathbf{X}_U$ . The very same experiments have been performed with  $\mathcal{D}_{UNC}$  as well. Again, all models have been trained for 50 epochs on  $128 \times 128$ px sized patches cropped around MS lesions. The respective performances on the testing sets of both domains are reported in Table 3 and Table 4.

Table 3: Domain Adaptation experiments for  $\mathcal{D}_{CHB} \rightarrow \mathcal{D}_{MS}$ 

<b>Model</b>	<b>MSSEG-CHB</b>			<b>MS</b>		
	<b>DICE</b>	<b>DICE (<math>\mu \pm \sigma</math>)</b>	<b>AUPRC</b>	<b>DICE</b>	<b>DICE (<math>\mu \pm \sigma</math>)</b>	<b>AUPRC</b>
$\mathbf{B}_{\mathcal{Y}_L}$	0.4473	$0.4472 \pm 0.0003$	0.3649	0.3975	$0.3752 \pm 0.0769$	0.3185
$\mathbf{B}_{\mathcal{Y}_L+\mathcal{S}}$	0.5756	$0.5423 \pm 0.0580$	0.5843	0.6751	$0.6547 \pm 0.0802$	0.6927
$\mathbf{B}_{\mathcal{Y}_L+\mathcal{Y}_U}$	0.5590	$0.5278 \pm 0.0580$	0.54081	0.7203	$0.6935 \pm 0.0646$	0.7597

Table 4: Domain Adaptation experiments for  $\mathcal{D}_{UNC} \rightarrow \mathcal{D}_{MS}$ 

<b>Model</b>	<b>MSSEG-UNC</b>			<b>MS</b>		
	<b>DICE</b>	<b>DICE (<math>\mu \pm \sigma</math>)</b>	<b>AUPRC</b>	<b>DICE</b>	<b>DICE (<math>\mu \pm \sigma</math>)</b>	<b>AUPRC</b>
$\mathbf{B}_{\mathcal{Y}_L}$	0.3924	$0.3903 \pm 0.0047$	0.3170	0.3622	$0.3428 \pm 0.0698$	0.3059
$\mathbf{B}_{\mathcal{Y}_L+\mathcal{S}}$	0.5634	$0.5314 \pm 0.0622$	0.5649	0.6746	$0.6611 \pm 0.0780$	0.6905
$\mathbf{B}_{\mathcal{Y}_L+\mathcal{Y}_U}$	0.5877	$0.5628 \pm 0.0467$	0.5804	0.7195	$0.6945 \pm 0.0708$	0.7433

### 3.5. Discussion

As Table 2 shows, training a supervised model ( $\mathbf{A}_{\mathcal{S}}$ ) only from artificial ground-truth performs considerably better than the UAD which actually produces this artificial ground-truth  $\mathcal{S}$ . We believe this occurs due to the fact that the supervised model is trained for an explicit objective, whereas the UAD approach has no knowledge about the task at hand. In comparison to the supervised model  $\mathbf{A}_{\mathcal{Y}_L}$  and the semi-supervised model  $\mathbf{A}_{\mathcal{Y}_L+\mathcal{S}}$ , we notice that  $\mathbf{A}_{\mathcal{S}}$  is slightly inferior, which might be due to FPs in segmentations (see Figure 3 C) provided by the UAD for training the segmentation network. The FP in  $\mathcal{S}$  are possibly learned and again reflected by the segmentation model. This effect might be overcome or at least weakened when using the continuous UAD output rather than binarizations. Interestingly, when using both  $(\mathcal{X}_L, \mathcal{Y}_L)$  and  $(\mathcal{X}_U, \mathcal{S})$  to train the semi-supervised model  $\mathbf{A}_{\mathcal{Y}_L+\mathcal{S}}$ , the resulting network is also inferior to the supervised model  $\mathbf{A}_{\mathcal{Y}_L}$ , although additional data has been provided. Again, we amount this to FP in  $\mathcal{S}$ . In fact, training data from 15 subjects seems to provide enough information for obtaining a model which performs already well, such that there might hardly be additional information in UAD delineations for the model to exploit. Simultaneously, the imperfections in  $\mathcal{S}$  might be learned, though, and potentially confuse the model.

For the task of Domain Adaptation, however, our framework shows great potential. A model  $\mathbf{B}_{\mathcal{H}}$  originally trained from 6 labeled subjects coming from dataset  $\mathcal{D}_{CHB}$  generalizes poorly to testing data from domain  $\mathcal{D}_{MS}$ , but when leveraging the artificial ground-truth  $\mathcal{S}$  provided by the UAD and also training the segmentation network from the originally unlabeled data, we witness great improvements on both domains. On the source domain  $\mathcal{D}_{CHB}$ , we even outperform the upper bound model  $\mathbf{B}_{\mathcal{H}_L + \mathcal{H}_U}$  which has been trained from labeled data of both domains. Similarly, although not outperforming the upper bound model, a positive trend is also noticed in the experiments involving the  $\mathcal{D}_{UNC}$  dataset, which provides clear evidence that UAD delineations can be very beneficial for Domain Adaptation.

## 4. Conclusion

We presented a novel framework which combines unsupervised deep representation learning and supervised deep learning into a pipeline which can be used for both Semi-supervised and completely Unsupervised Deep Learning. We believe that this approach can be useful beyond the presented use case of WML segmentation in brain MR, as long as the unsupervised anomaly detection provides labels at reasonable quality. In future work, we would like make use of the continuous UAD output rather than the binarized detections to make sure lower confidence anomalies have proportionally less impact on the segmentation performance of the supervised model. It might also be beneficial to follow the idea in (Dong et al., 2018) and employ a discriminator on the supervised segmentation network to regularize the model by encouraging good, realistic segmentations.

## Acknowledgments

We thank our clinical partners from Klinikum Rechts der Isar for generously providing us with their dataset.

## References

- Christoph Baur, Shadi Albarqouni, and Nassir Navab. Semi-supervised deep learning for fully convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 311–319. Springer, 2017.
- Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. *arXiv preprint arXiv:1804.04488*, 2018.
- Tom Brosch, Lisa YW Tang, Youngjin Yoo, David KB Li, Anthony Traboulsee, and Roger Tam. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.
- Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.

- Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018.
- Nanqing Dong, Michael Kampffmeyer, Xiaodan Liang, Zeya Wang, Wei Dai, and Eric Xing. Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 544–552. Springer, 2018.
- Pierre-Antoine Ganaye, Michaël Sdika, and Hugues Benoit-Cattin. Semi-supervised learning for segmentation under semantic constraint. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 595–602. Springer, 2018.
- J E Iglesias, Cheng-Yi Liu, P M Thompson, and Zhuowen Tu. Robust Brain Extraction Across Datasets and Comparison With Publicly Available Methods. *IEEE Transactions on Medical Imaging*, 30(9):1617–1634, 2011.
- Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Gig S Mageras, Joseph O Deasy, and Harini Veeraraghavan. Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 777–785. Springer, 2018.
- Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging*, pages 597–609. Springer, 2017.
- Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013.
- Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.
- Nick Pawlowski, Matthew CH Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan Stevenson, Aneesh Khetani, Tom Newman, et al. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders. In *International Conference on Medical Imaging with Deep Learnin (MIDL 2018)*, 2018.
- Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554, 2015.
- Torsten Rohlfing, Natalie M Zahr, Edith V Sullivan, and Adolf Pfefferbaum. The SRI24 multi-channel atlas of normal adult human brain structure. *Human Brain Mapping*, 31(5):798–819, December 2009.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Snehashis Roy, John A Butman, Daniel S Reich, Peter A Calabresi, and Dzung L Pham. Multiple sclerosis lesion segmentation from brain mri via fully convolutional neural networks. *arXiv preprint arXiv:1803.09172*, 2018.

Daisuke Sato, Shouhei Hanaoka, Yukihiro Nomura, Tomomi Takenaga, Soichiro Miki, Takeharu Yoshikawa, Naoto Hayashi, and Osamu Abe. A primitive study on unsupervised anomaly detection with an autoencoder in emergency head ct volumes. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105751P. International Society for Optics and Photonics, 2018.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.

Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramío-Torrenta, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.

# Learning interpretable multi-modal features for alignment with supervised iterative descent

**Max Blendowski**

**Mattias P. Heinrich**

*Institute of Medical Informatics, University of Lübeck, DE*

BLENDOWSKI@IMI.UNI-LUEBECK.DE

HEINRICH@IMI.UNI-LUEBECK.DE

## Abstract

Methods for deep learning based medical image registration have only recently approached the quality of classical model-based image alignment. The dual challenge of both a very large trainable parameter space and often insufficient availability of expert supervised correspondence annotations has led to slower progress compared to other domains such as image segmentation. Yet, image registration could also more directly benefit from an iterative solution than segmentation. We therefore believe that significant improvements, in particular for multi-modal registration, can be achieved by disentangling appearance-based feature learning and deformation estimation. In contrast to most previous approaches, our model does not require full deformation fields as supervision but rather only small incremental descent targets generated from organ labels during training. By mapping the complex appearance to a common feature space in which update steps of a first-order Taylor approximation (akin to a regularised Demons iteration) match the supervised descent direction, we can train a CNN-model that learns interpretable modality invariant features. Our experimental results demonstrate that these features can be plugged into conventional iterative optimisers and are more robust than state-of-the-art hand-crafted features for aligning MRI and CT images.

**Keywords:** Multi-Modal Features, Image Registration, Machine Learning.

## 1. Introduction

Much recent research has aimed at improving the alignment of images by means of learning optical flow (deformable registration) (Dosovitskiy et al., 2015; Hu et al., 2018). Deep convolutional networks for displacement field prediction share some similarities to conventional alignment strategies, but in general require a large amount of trainable parameters in a succession of convolution layers, which make the interpretation of learned features difficult. In addition, most previous work has focused on image sequences of the same modality with only subtle changes in appearance and lighting. However, medical image analysis in particular requires the comparison of related structures in different modalities. Here, the importance lies in obtaining interpretable cross-modal features that enable meaningful correspondences for evaluating changes across patients, to aid diagnosis and for biomarker discovery.

**Related Work:** In general, most classical multi-modal image registration approaches make use of two core ideas: either they rely on a similarity measure that is able to handle multi-modal input or they try to find a mapping for both modalities to a shared space and use a monomodal metric. Based on information theoretic insights and as a representative of the first group, (Maes et al., 1997) introduced Mutual Information as a similarity measure that does not require cross-modal features. Exemplary for the second group, inspired by the concept of Self Similarity proposed in (Shechtman and Irani, 2007), (Heinrich et al., 2012) introduced the expressive, cross-modal MIND

descriptor that allows the usage of standard similarity metrics, e.g. the sum of squared differences. (Kim et al., 2017) present end-to-end trainable CNN-based self similarity features, however only trained on mono-modal input. Modality conversion has been employed for learning transferable representations with unpaired multi-modal CycleGANs (Tanner et al., 2018) and (Mahapatra et al., 2018) use GANs for multimodal image registration.

A variety of recent learning based registration methods has emerged that, in contrast to classical modular techniques, comprise the whole process to generate a displacement field from a given image pair in a fully integrated feed-forward step. Therefore, it is difficult to determine which parts are responsible for alignment or feature extraction in methods that resemble fully-convolutional encoder-decoder architectures, e.g the SVF-Net (Rohé et al., 2017) or (Balakrishnan et al., 2018). Furthermore, due to their high number of parameters, e.g. the FlowNet proposed by (Dosovitskiy et al., 2015) or the label-driven, weakly supervised method of (Hu et al., 2018) require large datasets with (pseudo-)ground truth labels during training. We take inspiration from recent work in computer vision, where (Brachmann et al., 2017) developed DSAC (differentiable RANSAC) as a modular end-to-end trainable fitting approach, which effectively disentangles feature learning from regressing a transformation - but so far only for low-parametric homographies.

**Overview and contributions:** Our strategy is to integrate deep learning methods into the classical registration pipeline by learning expressive cross-modal features. Similar to (Xiong and De la Torre, 2013) with their supervised descent approach, we also learn features for a descent direction. However, their method is restricted to mono-modal data and population-based, thus inapt for one-to-one alignment problems. Based on regression forests, (Gutierrez-Becker et al., 2017) learn guided update steps by recombining an ensemble of input features. Instead, we aim to learn these features from scratch and give a detailed explanation of our approach in the following. In contrast to current work on weakly-supervised optical flow learning, we focus on the disentanglement of appearance and deformation (which has also seen interest in face analysis (Shu et al., 2018)). We thus propose a new **S**Upervised **I**Terative de**S**cent algorithm (SUTS) that unrolls a conventional iterative optimisation of regularised B-spline transformations into a differentiable (recurrent) network. By employing the generally applicable constraints of regularised iterative alignment, our model requires only very few trainable weights for learning expressive multi-modal features and can be trained with small datasets under only weak-supervision of segmentation labels. Our experimental validation on multi-modal CT to MRI registration achieves encouraging improvements over hand-crafted features and serves as proof of concept.

## 2. Methods

In this section, we introduce our proposed supervised iterative descent algorithm (SUTS) for multi-modal image registration. Figure 1 shows the general structure of the method for both the training and inference phases. In addition, the modular interrelationships are indicated, which allow the learning of meaningful feature extracting networks and an iterative estimation of the displacement fields. First, the differentiable B-Spline Descent module essential for this approach is presented, before the entire process is explained in more detail.

### 2.1. B-Spline Descent module

The main source of inspiration for our work is the classical pipeline of feature-based iterative image registration. If the images  $f$  (fixed) and  $m$  (moving) are displayed in a common characteristic space

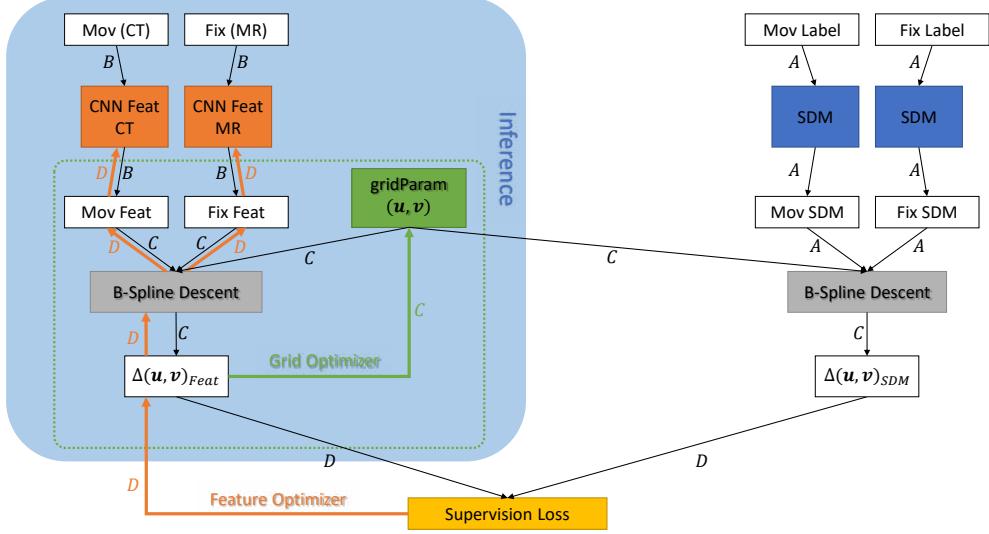


Figure 1: Method overview: During training our approach employs two Adam optimizers; *Grid Optimizer* (green) updates the displacement field parameters based on incremental steps  $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$  per B-Spline Descent module iteration ( $C$ ). *Feature Optimizer* (orange) adapts the feature CNNs’ weights (one complete run  $B-D$ ) guided by the supervision signal stemming from differences to the SDM ( $A$  - precomputed once) based incremental steps. During inference, only *Grid Optimizer* adjusts the displacement grid, operating on fixed features (MIND/CNNFeat).

by the selection of suitable features, it allows the use of optical flow methods, because the brightness consistency constraint, which results in a sum of squared differences metric, is approximately valid. Based on this assumption, we introduce a B-Spline descent module (grey B-Spline Descent Block in Figure 1). Multi-channel feature representations  $M$  and  $F$  (per pixel 8 channels for the CNN-based approach and 6 for MIND descriptors) of  $m$  and  $f$  as well as their current displacement parameters  $(\mathbf{u}, \mathbf{v})$  serve as input to compute incremental updates  $\Delta(\mathbf{u}, \mathbf{v})$  for the displacements as output.  $(\mathbf{u}, \mathbf{v})$  holds a two-dimensional displacement vector at every pixel and  $\Delta(\mathbf{u}, \mathbf{v})$  contains the gradients to update the displacements. We adopt a widely used energy term to compute  $\Delta(\mathbf{u}, \mathbf{v})$ . To simplify their calculation, a linearization using a first order Taylor approximation is performed - only valid under the assumption of small displacement updates per step used to iteratively refine the warping result based on the initial moving image (Papenberg et al., 2006). Per *image channel*  $c$  and pixel position this leads to

$$E_c(\mathbf{u}_c(\mathbf{x}), \mathbf{v}_c(\mathbf{x})) = \frac{1}{2} (M_c(\mathbf{x}) + M_{c,\partial x} \cdot \mathbf{u}_c(\mathbf{x}) + M_{c,\partial y} \cdot \mathbf{v}_c(\mathbf{x}) - F_c(\mathbf{x}))^2 + \frac{\lambda}{2} (\mathbf{u}_c(\mathbf{x}) + \mathbf{v}_c(\mathbf{x}))^2 \quad (1)$$

Here,  $M_{c,\partial x/y}$  denote the partial moving image derivatives of channel  $c$  and  $\frac{\lambda}{2} (\mathbf{u}_c(\mathbf{x}) + \mathbf{v}_c(\mathbf{x}))^2$  penalizes big displacement updates. Taking the partial derivatives  $\frac{\partial E_c(\mathbf{u}_c, \mathbf{v}_c)}{\partial \mathbf{u}_c(\mathbf{x})/\mathbf{v}_c(\mathbf{x})}$  to minimize this expression

and sorting the terms into a linear system of equations yields:

$$\begin{bmatrix} M_{c,\partial_x}^2 + \lambda & M_{c,\partial_x} M_{c,\partial_y} \\ M_{c,\partial_x} M_{c,\partial_y} & M_{c,\partial_y}^2 + \lambda \end{bmatrix} \begin{bmatrix} \mathbf{u}_c(\mathbf{x}) \\ \mathbf{v}_c(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} (F_c - M_c) M_{c,\partial_x} \\ (F_c - M_c) M_{c,\partial_y} \end{bmatrix} \quad (2)$$

Finally, using the Sherman-Morrison-Woodbury formula a matrix inversion-free expression is used to efficiently compute displacement grid updates:

$$\begin{bmatrix} \mathbf{u}_c(\mathbf{x}) \\ \mathbf{v}_c(\mathbf{x}) \end{bmatrix} = \frac{1}{\lambda + M_{c,\partial_x}^2 + M_{c,\partial_y}^2} \cdot \begin{bmatrix} (F_c - M_c) M_{c,\partial_x} \\ (F_c - M_c) M_{c,\partial_y} \end{bmatrix} \quad (3)$$

Following the heuristic presented in (Guimond et al., 2002), we solve (2) using (3) independently per channel and average the individual, channel-wise solutions to obtain the displacement updates  $\Delta(\mathbf{u}, \mathbf{v})$ . This update step is basically the core of our approach. It is composed only of differentiable operations and therefore easily to integrate into backpropagation engines used for machine learning such as PyTorch.

In accordance with most image registration approaches, our module also allows the use of displacement fields with a more coarse grid spacing with respect to the actual image. In order to generate dense fields to warp the moving image, we employ third order cardinal B-Spline interpolations (Tustison and Avants, 2013). Due to their definition on a uniform spaced grid and their recursive formulation, interpolation between the knots equals a convolution operation with a smoothing kernel. Using differentiable upsampling followed by two average pooling layers, we can efficiently implement this interpolation scheme to generate dense displacement fields. While choosing  $\lambda = (M - F)^2$  locally adapting following (Vercauteren et al., 2009), the incremental displacement updates often exhibit implausible strong local changes. On that account, an additional smoothness penalty that considers the deviation of  $\Delta(\mathbf{u}, \mathbf{v})$  from a smoothed version of itself is introduced.

Overall, it is worth noting, that our B-Spline Descent module already outputs an update direction with  $\Delta(\mathbf{u}, \mathbf{v})$  that can be backpropagated by off-the-shelf optimizers. Details on how to use the module in the larger context of our method follow below.

## 2.2. Supervised Iterative Descent (SUTS)

With the B-Spline Descent module at hand, image pairs fulfilling the brightness consistency constraint, should be able to be aligned with each other by iteratively updating the displacement grid parameters. However, our goal is to align multi-modal images that violate this constraint. To this end, we want to learn CNN-based feature mappings for both modalities from scratch - each representing a mapping to a shared representation. This raises the question of how to learn these feature mappings.

**Learning Feature CNNs:** Here, our idea is to use a form of weak supervision in order to achieve a meaningful gradient guidance for error backpropagation to train the feature CNNs. As depicted in Figure 1, during training we make use of an auxiliary image representation that is computed and processed in the right-hand stream: while using unpaired multi-modal inputs, for each individual image there are organ segmentations available, which we transform to their signed distance maps (SDM) to represent a simple form of a shared feature space (A). We assume that inserting SDMs into the B-Spline Descent module will yield  $\Delta(\mathbf{u}, \mathbf{v})_{SDM}$  as a sufficient guidance for a single descent step of  $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$  - even if the SDM representation is incomparable (directly) to the original image input.

In our case of unpaired image data, we depend only on this supervisory gradient signal that can be used to propagate update information to the feature CNN weights. An Adam optimizer (named *Feature Optimizer*), that keeps track of which operations the image pairs undergo until  $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$  is generated, performs updates only on the weights of the feature CNNs ( $D$ ). The loss signal for this backpropagation is simply the mean squared error  $MSE(\Delta(\mathbf{u}, \mathbf{v})_{Feat}, \Delta(\mathbf{u}, \mathbf{v})_{SDM})$ .

**Iterative Image Alignment:** In order to iteratively align the input images, we employ a second Adam optimizer (named *Grid Optimizer*). The green dotted box in Figure 1 encloses the region where it is active. Based on their alignment due to the displacements  $(\mathbf{u}, \mathbf{v})$  derived from the current (learned) feature representations of two images ( $B$ ), the B-Spline Descent module outputs an descent step  $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$  ( $C$ ). This is fed into the *Grid Optimizer* as gradient value for the parameters  $(\mathbf{u}, \mathbf{v})$  when it performs its update step. Note that these parameters  $(\mathbf{u}, \mathbf{v})$  are shared for both the CNN- and the SDM-stream of our method. Also, they are only indirectly influenced by the SDM stream through the supervision signal during training when updating the feature CNNs' weights ( $D$ ) (this means we never directly use an SDM gradient for spatial alignment). Since the displacement grid parameters  $(\mathbf{u}, \mathbf{v})$  and their momentum act as memory encoding the current alignment state when adjusting the feature weights, our approach can be interpreted as a recurrent method. Nevertheless, to generate robust features that have to remain fixed during inference, it is important to consider image pairs of different alignment stages throughout training. We achieve this by dividing a given number of total iterations per image pair into  $p$  subintervals and randomly choosing the pair to be presented at each step. Thereby within one mini-batch image pairs with different degrees of alignment (number of current update steps) will be used at the same time. Here, we differ from multi-stage regression approaches as proposed e.g. in (Xiong and De la Torre, 2013), since we want our learned features (that are fixed during inference) to be expressive and applicable during the *complete* iterative alignment process for a given image pair.

Our method allows to use multiscale strategies by a stepwise refinement of the grid spacing: starting from a coarse control point grid and performing a fixed number of incremental displacement updates, the displacement field parameters  $(\mathbf{u}, \mathbf{v})$  of the next stage with a smaller spacing are initialized by upscaled versions of their predecessors.

During inference, the CNN feature representations of an unseen image pair - now without the need of additional annotations - will be iteratively aligned for a fixed number of iterations at different gridscales (indicated by the blue box). Finally, their resulting displacement parameters  $(\mathbf{u}, \mathbf{v})$  (green block) can be used to warp the moving towards the fixed image. For further algorithmic details refer to Algorithm 1 in Appendix A.<sup>1</sup>

### 3. Experiments & Results

To verify the applicability of our approach, we perform multi-modal image registrations on 2D coronal slices of unpaired abdominal CT and MRI scans from the VISCERAL dataset (Jimenez-del Toro et al., 2016) with a similar slice thickness. As additional information to train our method, we use the provided label maps for the following structures: liver, spleen, kidneys and psoas major muscles. Dice scores that our approach achieves with fixed learned features during testing serve as our quality measure. We resample these images to an isotropic pixel size of  $1.5\text{mm}^2$  and compensate only for through-plane transformations of the 3D volumes using the deeds-SSC approach of (Heinrich et al., 2013) and leaving all non-rigid in-plane deformations, which results in a large

---

1. We plan to make our code publicly under [https://github.com/multimodallearning/midl19\\_suits](https://github.com/multimodallearning/midl19_suits).

initial misalignment with Dice overlap of 0.44. Subsequently, we crop them without any guidance to dimensions of 320x312. Figure 2 shows example slices from this dataset along with signed distance maps for different organ structures. First, we conducted an unsupervised experiment for monomodal

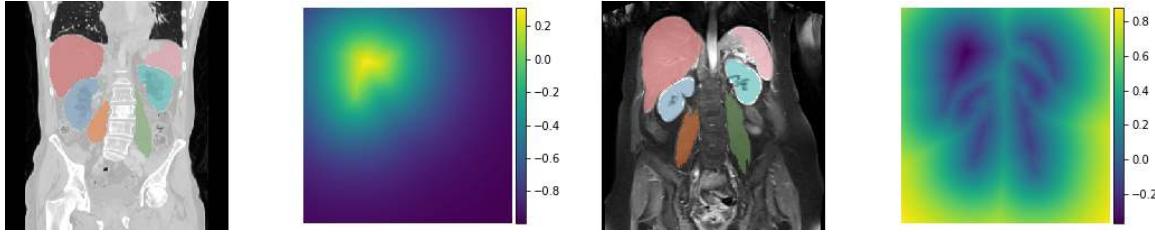


Figure 2: Exemplary Abdominal Scans: (left to right) CT image with provided expert segmentations; scaled signed distance map (SDM) of the liver (CT); MR image with provided expert segmentations; SDM of the background (MR)

CT registration on image intensities to demonstrate the suitability of our B-Spline Descent module for general purposes. Here, initial mean Dice scores increased from 0.44 to 0.69.

In order to assess the results of our proposed *multi-modal* approach, we use the non-trainable MIND descriptor (Heinrich et al., 2012) within our B-Spline Descent scheme as comparison that has been shown to provide robust and expressive modality-independent features for image registration. We also conduct experiments with a state-of-the-art algorithm (SimpleElastix) for multi-modal images (metric: mutual information, 4 level multi-resolution alignment with affine pre-registration) as additional baseline.

In total, we use 10 slices per modality, which corresponds to 100 possible registration pairs. To train and evaluate our approach, we split our patient set in 7 randomly drawn training slices per modality (49 pairs) and evaluate the registration accuracy on the remaining 9 pairs. We repeat this procedure 10 times and provide the mean Dice scores per organ structure for both the MIND baseline as well as for our learned features.

As feature extractors, we use two identical architectures for both modalities: a simple feed-forward design consisting of 7 convolutional blocks each. Using appropriate padding to maintain the input size, we encode the characteristics of the convolutional layers as tuples containing (kernel\_size, #output\_channels). The number of input channels of the first layer equals 1, since it processes the original images. The following layers choose their number of input channels according to the output channels of their predecessor. Thus, the sequence (7,4)-(7,6)-(5,6)-(5,8)-(5,8)-(3,8)-(3,8) unambiguously defines our feature learning networks. Except from the last convolution that also learns its bias values, they are followed by Group normalization blocks (2 groups) and *tanh*-activation functions. During training, the SDMs are used to supervise the gradient signal. They are computed for the background as well as for every organ based on the expert segmentations and stored channelwise. Since the raw values of the SDMs exhibit large variations depending on different sizes or positions for each organ, we normalize their range to  $[-1, 1]$  by applying a  $\tanh(\alpha \cdot x)$  function, where the scaling factor  $\alpha$  is set to 0.01 (see Figure 2 for examples). Due to this, we ensure similar value ranges for the SDM feature maps and the CNN feature outputs.

Taking the training schedule for robust feature learning into account as described above (preventing the network from seeing only already similarly aligned images), we update the feature CNNs' weights with *Feature Optimizer* after every 5th iteration of *Grid Optimizer*. *Feature Optimizer* uses an initial learning rate of 0.001 for its parameter updates, while using 0.005 for *Grid Optimizer*. In total, we use 3 grid spacing scales, refining the control point spacing from 20, over 10 to every 7th pixel position. At each stage we compute 300 updates  $\Delta(\mathbf{u}, \mathbf{v})$  for the displacement grid parameters. The additional penalty, that considers deviations of  $\Delta(\mathbf{u}, \mathbf{v})$  from its smoothed version, is given more weighting in the finest stage with 0.025 compared to 0.0125 at the first ones. We use two image pairs as one batch. Finally, we only backpropagate updates from regions around organ borders with *Feature Optimizer* to learn CNN weights, i.e. where  $\text{abs}(\text{SDM}_{f/m}(\mathbf{x})) < 0.1$ , because only in this regions the supervision can be expected to be informative.

At test time, we maintain all parameters as described above, except that our Feature CNNs are fixed now, i.e. only *Grid Optimizer* performs its iterations. Also, when extracting features with the MIND descriptor instead of our trained CNN features, we keep these parameters. There is no longer a need for the SDM stream as depicted in Figure 1, since it is only used for the supervisory signal during training.

**Results:** Qualitative results are illustrated in Figure 3 for CT to MR registrations. Here, for both descriptor extraction methods results are depicted. The top row shows the initial images overlayed with their respective segmentations, followed by the fixed image overlayed with segmentations of the moving image that have been warped according to the displacement fields generated based on MIND- and CNN-based features (MIND-Dice: 0.63, CNN-Dice: 0.74), respectively. The bottom row first illustrates the CNN-based displacements by its effects on a grid image after all stages and iterations with the B-Spline Descent module. The following checkerboard visualization gives an intuition of the initial organ alignment (Mean Dice Score 0.40). Also, for both feature types, the checkerboard images depict the organ alignment after registration, where the approach achieves Dice scores of 0.65 based on MIND-features and 0.75 with the learned features for this exemplary image pair.

More quantitative results can be found in Table 1. Here, the mean after 10 test iterations (a total of 90 registration pairs) using the CNN based features slightly outperforms the MIND features. While the MIND experiments achieves better results for the psoas muscles, the trained features (Mean Dice Score 0.72) yield better results aligning larger organ structures, such as the liver or the spleen. Moreover, it compares favourably well against SimpleElastix (Mean Dice Score 0.70) as representative of elaborate, classical algorithms.

#### 4. Discussion & Conclusion

In our work, we developed a new approach to integrate CNN-based multi-modal features into the classical image registration pipeline. We showed that even with unpaired images and only a weak supervision by few organ labels during training, expressive image representations in a shared space are generated - readily employable in iterative multi-stage alignment frameworks. While especially for smaller structures handcrafted MIND features perform competitively, due to the increased receptive field of a multi-layer CNN, larger structures can more easily be aligned. Overall, our method enables to replace the need for dense correspondences by costly to generate displacement fields with organ label maps as the only required additional information during training.

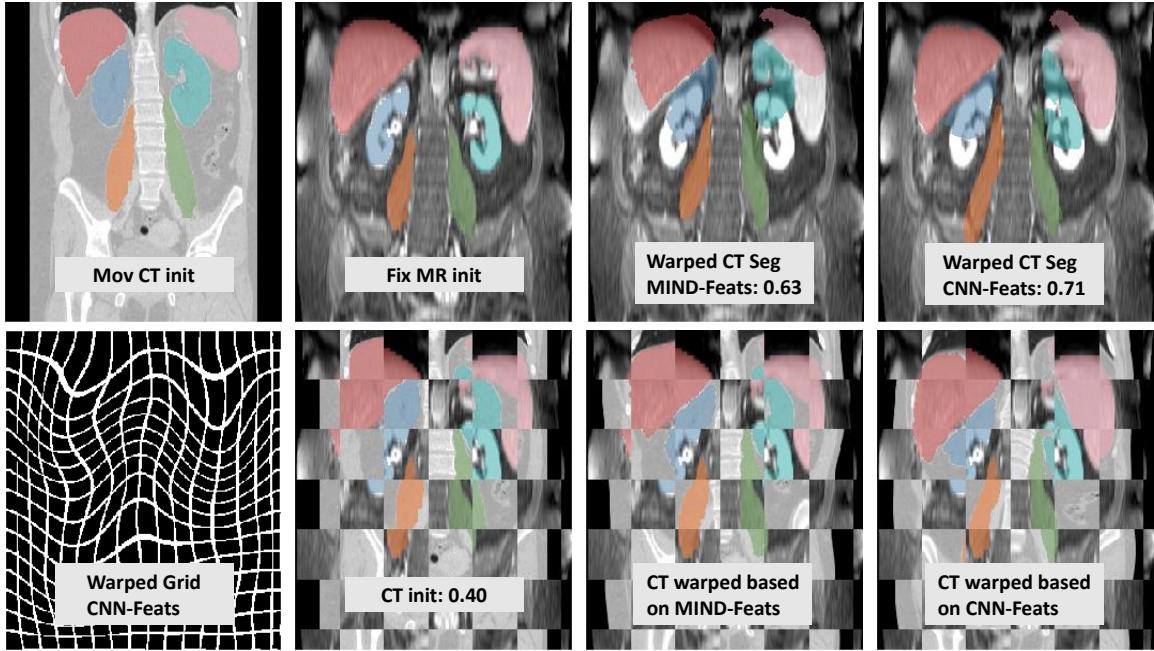


Figure 3: Exemplary Results: (Top row) CT & MRI scans overlaid with respective segmentations and illustration of the MRI scan overlaid with the warped segmentations based on MIND (Dice 0.63) and learned CNN features (Dice 0.71). (Bottom row) An accordingly to the CNN feature based displacement field warped grid; checkerboard illustrations of the initial organ misalignments (Dice: 0.40) and after warping with the respective displacement fields.

Table 1: Organ Dice Scores Listings: Compared to the initial overlaps, both feature extraction approaches achieve better alignments. While the MIND-based registrations perform competitively especially on fine structures (Psoas muscles), the CNN-Feature-based alignments yield favourable results in particular for large structures (liver, spleen).

<b>Experiment</b>	<b>Organs</b>							$\emptyset$
	Liver	Spleen	LKidney	RKidney	LPsoas	RPsoas		
Initial Overlap	0.56	0.37	0.52	0.55	0.53	0.65		0.53
SimpleElastix	0.75	<b>0.68</b>	0.58	<b>0.72</b>	0.68	<b>0.76</b>		0.70
MIND Descriptor	0.67	0.45	0.70	0.69	<b>0.72</b>	0.75		0.66
Feature CNNs	<b>0.83</b>	0.64	<b>0.74</b>	0.68	<b>0.72</b>	0.73		<b>0.72</b>

For future work, this proof of concept encourages to investigate our method in several ways. First, with an extension to 3D, our approach can be examined on challenging 3D datasets with higher demands on memory and computational restrictions. Also, an evaluation of effects depending on

architectural design choices regarding the feature generating networks clearly is of interest. Finally, studying the effects of other possible supervision signals, e.g. combining SDM and MIND features as auxiliary guidance, will provide further insights.

To conclude with, our experiments support the assumption, that disentangling appearance-based feature learning and deformation estimation - as practised in traditional, well-studied iterative approaches - can provide an alternative to parameter-intense end-to-end CNN registration methods.

## Acknowledgments

This work was supported by the German Research Foundation (DFG) under grant number 320997906 (HE 7364/2-1). We would like to thank the reviewers for their many insightful comments and suggestions helping to improve our paper. We gratefully acknowledge the support of the NVIDIA Corporation with their GPU donations for this research.

## References

- Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: A learning framework for deformable medical image registration. *arXiv preprint arXiv:1809.05231*, 2018.
- Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 2017.
- Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.
- Alexandre Guimond, Charles RG Guttmann, Simon K Warfield, and C-F Westin. Deformable registration of dt-mri data based on transformation invariant tensor characteristics. In *Proceedings IEEE International Symposium on Biomedical Imaging*, pages 761–764. IEEE, 2002.
- Benjamin Gutierrez-Becker, Diana Mateus, Loic Peter, and Nassir Navab. Guiding multimodal registration with learned optimization updates. *Medical image analysis*, 41:2–17, 2017.
- Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical image analysis*, 16(7):1423–1435, 2012.
- Mattias P Heinrich, Mark Jenkinson, Michael Brady, and Julia A Schnabel. Mrf-based deformable registration and ventilation estimation of lung ct. *IEEE transactions on medical imaging*, 32(7):1239–1248, 2013.
- Yipeng Hu, Marc Modat, Eli Gibson, Wenqi Li, Nooshin Ghavami, Ester Bonmati, Guotai Wang, Steven Bandula, Caroline M Moore, Mark Emberton, et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis*, 49:1–13, 2018.

- Oscar Jimenez-del Toro, Henning Müller, Markus Krenn, Katharina Gruenberg, Abdel Aziz Taha, Marianne Winterstein, Ivan Eggel, Antonio Foncubierta-Rodríguez, Orcun Goksel, András Jakab, et al. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. *IEEE transactions on medical imaging*, 35(11):2459–2475, 2016.
- Seungryong Kim, Dongbo Min, Bumsuk Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *Proc. IEEE Conf. Comp. Vision Patt. Recog*, page 8, 2017.
- Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.
- Dwarikanath Mahapatra, Bhavna Antony, Suman Sedai, and Rahil Garnavi. Deformable medical image registration using generative adversarial networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1449–1453. IEEE, 2018.
- Nils Papenberg, Andrés Bruhn, Thomas Brox, Stephan Didas, and Joachim Weickert. Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision*, 67(2):141–158, 2006.
- Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svfnet: Learning deformable image registration using shape matching. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 266–274. Springer, 2017.
- Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*, pages 1–8. IEEE, 2007.
- Zhixin Shu, Mihir Sahasrabudhe, Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. *arXiv preprint arXiv:1806.06503*, 2018.
- Christine Tanner, Firat Ozdemir, Romy Profanter, Valeriy Vishnevsky, Ender Konukoglu, and Orcun Goksel. Generative adversarial networks for mr-ct deformable image registration. *arXiv preprint arXiv:1807.07349*, 2018.
- Nicholas James Tustison and Brian Avants. Explicit b-spline regularization in diffeomorphic image registration. *Frontiers in neuroinformatics*, 7:39, 2013.
- Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.
- Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

## Appendix A. Algorithmic Details

---

**Algorithm 1:** Schematic overview of the Training procedure

---

**Input:** CT & MR images + organ labels  
**Output:** CNNs trained for Feature Extraction

Initialize FEATURE CNNs;  
 Initialize FEATURE OPTIMIZER & register the CNNs WEIGHTS;  
 Initialize GRID OPTIMIZER & register the displacement parameters  $(\mathbf{u}, \mathbf{v})$ ;  
 Generate PAIRPRESENTATIONSCHEME; // different alignment stages  
 Compute Fixed Signed Distance Maps  $M_{SDM}$  &  $F_{SDM}$ ; // cf. As in Figure 1  
**for** #grid\_scales **do**  
   **while** batch\_pairs **in** PAIRPRESENTATIONSCHEME **do**  
     // Tracked by FEATURE OPTIMIZER  
     Compute  $M_{feat} = \text{CNN}_{CT}(m)$  &  $F_{feat} = \text{CNN}_{MRI}(f)$ ; // cf. Bs  
     // NOT Tracked by FEATURE OPTIMIZER  
     **for** #grid\_iters **do**  
       // Perform several displacement grid update steps  
       Compute GridUpdate  $\Delta(\mathbf{u}, \mathbf{v})_{Feat} = \text{BSTModule}(M_{Feat}, F_{Feat}, (\mathbf{u}, \mathbf{v}))$ ; // cf. Cs  
       Use GRID OPTIMIZER to update  $(\mathbf{u}, \mathbf{v})$  by  $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$ ;  
     **end**  
     // Tracked by FEATURE OPTIMIZER  
     Compute  $\Delta(\mathbf{u}, \mathbf{v})_{Feat} = \text{BSTModule}(M_{Feat}, F_{Feat}, (\mathbf{u}, \mathbf{v}))$ ;  
     Compute  $\Delta(\mathbf{u}, \mathbf{v})_{SDM} = \text{BSTModule}(M_{SDM}, F_{SDM}, (\mathbf{u}, \mathbf{v}))$ ;  
     Compute MSE( $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$ ,  $\Delta(\mathbf{u}, \mathbf{v})_{SDM}$ ) as loss; // cf. Ds  
     Use FEATURE OPTIMIZER to update the CNN WEIGHTS  
**end**  
**end**

---



---

**Algorithm 2:** Schematic overview of the Pairwise Registration during Inference

---

**Input:** CT & MR image pairs; CNNs trained by Alg. 1  
**Output:** Warped Moving Image, Displacement Parameters  $(\mathbf{u}, \mathbf{v})$

Initialize GRID OPTIMIZER & register the displacement parameters  $(\mathbf{u}, \mathbf{v})$ ;  
**for** #grid\_scales **do**  
   Compute  $M_{feat} = \text{CNN}_{CT}(m)$  &  $F_{feat} = \text{CNN}_{MRI}(f)$ ; // cf. Bs  
   **for** #iters\_per\_scale **do**  
     Compute GridUpdate  $\Delta(\mathbf{u}, \mathbf{v})_{Feat} = \text{BSTModule}(M_{Feat}, F_{Feat}, (\mathbf{u}, \mathbf{v}))$ ; // cf. Cs  
     Use GRID OPTIMIZER to update  $(\mathbf{u}, \mathbf{v})$  by  $\Delta(\mathbf{u}, \mathbf{v})_{Feat}$ ;  
   **end**  
**end**  
 Warp  $m$  according to  $(\mathbf{u}, \mathbf{v})$ ;  
**return** warpedMovingImage,  $(\mathbf{u}, \mathbf{v})$

---

# Learning from sparsely annotated data for semantic segmentation in histopathology images

**John-Melle Bokhorst<sup>1</sup>**

JOHN-MELLE.BOKHORST@RADBOUDUMC.NL

**Hans Pinckaers<sup>1</sup>**

HANS.PINCKAERS@RADBOUDUMC.NL

**Peter van Zwam<sup>2</sup>**

P.VANZWAM@PAMM.NL

**Iris Nagtegaal<sup>1</sup>**

IRIS.NAGTEGAAL@RADBOUDUMC.NL

**Jeroen van der Laak<sup>1</sup>**

JEROEN.VANDERLAAK@RADBOUDUMC.NL

**Francesco Ciompi<sup>1</sup>**

FRANCESCO.CIOMPI@RADBOUDUMC.NL

<sup>1</sup> *DIAG Nijmegen, Geert Grootplein Zuid 10, 6525 GA The Netherlands*

<sup>2</sup> *PAMM Laboratory for Pathology and Medical Microbiology, Eindhoven, The Netherland*

## Abstract

We investigate the problem of building convolutional networks for semantic segmentation in histopathology images when weak supervision in the form of sparse manual annotations is provided in the training set. We propose to address this problem by modifying the loss function in order to balance the contribution of each pixel of the input data. We introduce and compare two approaches of loss balancing when sparse annotations are provided, namely (1) instance based balancing and (2) mini-batch based balancing. We also consider a scenario of full supervision in the form of dense annotations, and compare the performance of using either sparse or dense annotations with the proposed balancing schemes. Finally, we show that using a bulk of sparse annotations and a small fraction of dense annotations allows to achieve performance comparable to full supervision.

**Keywords:** Weakly supervised semantic segmentation, loss balancing, partially labelled data, computational pathology.

## 1. Introduction

The ability of computers to extract information from images has increased tremendously since convolutional neural networks (CNNs) have been introduced. For multiple years now, CNNs have been successfully applied to classification and segmentation tasks. Segmentation in medical imaging is the process of delineating the boundaries of various structures or tissues. As an example, in histopathology images of colorectal cancer (CRC), distinguishing glands (both healthy and cancerous) from surrounding connecting tissue (i.e., stroma) can be the basis of prognostic biomarkers, such as the tumor-stroma ratio ([Mesker et al., 2007](#)) ([Geessink et al., 2019](#)).

In semantic segmentation, supervised training of models usually requires labor intensive pixel annotations, which consist in a *dense* segmentation map (Figure 1(a)). In this approach, all pixels, mostly within a pre-fixed area, are assigned to one class by a human annotator. In the field of medical imaging and in particular of histopathology, this approach is not only labor intensive but also requires specialist knowledge about the transition between the different tissue types. Dense annotations allow a model to learn the transition between different classes, which is expected to produce an accurate semantic segmentation output. This approach can be considered as full supervision of segmentation models.

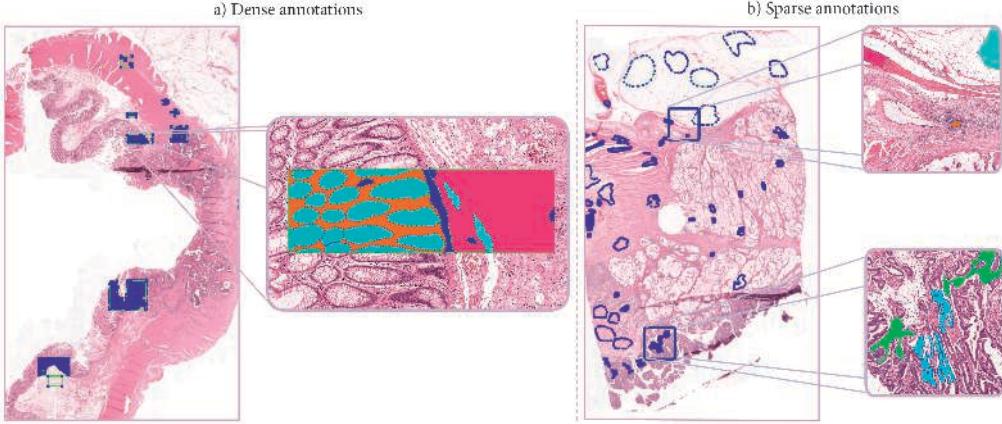


Figure 1: a) Example of a densely annotated image, b) Example of a sparsely annotated image.

An alternative to a fully supervised approach to segmentation is weak supervision, which can be provided in the form of bounding boxes, image level labels, dots or *sparse* (or partial) annotations (Figure 1(b)). In all these cases, only one or more parts of a tissue/class is labeled. In (Rajchl et al., 2017), bounding boxes were used for the segmentation of brain tissue, while (Ker-vadec et al., 2018) adapts the loss function for segmenting cardiac images with data annotated with scribbles. In (Glocker et al., 2013), a semi-automatic labeling strategy is proposed where sparse dot annotations are converted into dense probabilistic labels for vertebrae localization and naming. In histopathology images, (Xu et al., 2014) proposed to segment glands in CRC based on bounding boxes, however this method is not based on convolutional networks.

Sparse annotations allow to easily include more pixels than a scribble-based approach, but on the one hand do not guarantee to provide clear definition of transitions between different classes, and on the other hand provide pixel-level labels without the localization carried by bounding boxes. One typical use case of sparse annotations is focusing only on areas where the expert is absolutely certain about a specific class. Furthermore, it allows to quickly annotate a large variety of tissue types without having to focus on the surrounding of each specific class, which helps in the case of semantically under-represented tissues. For these reasons, sparse annotations may be considered as an attractive approach to create reference standard in medical imaging for building supervised segmentation models.

In this paper, we address the problem of multi-class semantic segmentation when *sparse* annotations are provided. For this purpose, we tackle the problem of class imbalance and lack of annotated pixels in training examples by modifying the loss function. We formulate two strategies to weigh the loss, namely (a) *instance based balancing* and (b) *mini-batch based balancing*. To investigate the proposed approach, we also consider a set of dense annotations and train segmentation models on fully annotated images alone as well as models that are given mainly sparsely annotated images and a few densely annotated images. We validate our approach on a tissue segmentation problem in colorectal cancer histopathology images, and we use U-Net as the CNN architecture for semantic segmentation. To the best of our knowledge we are the first to try semantic segmentation with sparsely annotated data in CRC histopathology images.

## 2. Materials

Seventy paraffin-embedded tissue specimens from colorectal cancer patients of the Radboud University Medical Center (Nijmegen, Netherlands) were included. Tissue slides were prepared and stained with H&E staining, and digitalized using a Pannoramic P250 Flash II scanner (3D-Histech, Hungary) at a spatial resolution of  $0.24 \mu\text{m}/\text{px}$ .

A pathologist and two trained human analysts were involved in manual annotations of whole-slide images. The set of cases was split into two parts and both sparse and dense annotations were made: sparse annotations were made on 54 images, dense annotations were made on 16 images. In order to make dense annotations, areas of different sizes, showing at least 2 tissue classes and the border area between them, were selected and annotated. In all images the following 13 tissue types where annotated; 1) tumor, 2) desmoplastic stroma, 3) necrosis and debris, 4) lymphocytes, 5) erythrocytes, 6) muscle, 7) healthy stroma, 8) fatty tissue, 9) mucus, 10) nerve, 11) stroma lamina propria, 12) healthy glands, 13) background. The ratios between the amount of annotated pixels per class for both datasets is shown in Table 1.

The set of whole-slide images (WSI's) with corresponding annotations was randomly divided into a training set (43 WSI's with sparse annotations, 8 WSI's with dense annotations), a validation set (11 WSI's with sparse annotations, 2 WSI's with dense annotations) and a test set, containing 5 WSI's with only dense annotations.

## 3. Method

When training a segmentation network like U-Net with mini-batch gradient descent (i.e., mini-batch size  $> 1$ ), attention should be paid to the contribution of individual pixels to the loss function. When sparse annotations are used it may occur that (1) not all classes are present equally in a mini-batch or within a patch and (2) not all pixels within the patch have been assigned to a label, as shown in Figure 2a. In order to tackle these problems, we investigate the effect of modifying the loss function based on the type of manual annotations of input training data. Inspired by the original work on the U-Net model, we define a weight map  $W$  that specifies the contribution of each pixel to the loss function  $L$ . In practice, if  $w_{ij}$  and  $l_{ij}$  are the weight map and the loss value for a pixel in position  $(i, j)$ , using a the weight map produces a new  $\hat{l}_{ij} = w_{ij}l_{ij}$  loss for each pixel. We introduce and compare two strategies to create such a weight map, based on different loss balancing strategies, namely (1) *instance based balancing*, and (2) *mini-batch based balancing*. We also compare these approaches with a case without balancing. These three approaches are formulated in detail in this section.

	Dense	Sparse
Tumor	17.43	4.47
Desmoplastic stroma	13.68	5.16
Necrosis and debris	1.36	9.80
Lymphocytes	1.80	3.83
Erythrocytes	0.67	2.34
Muscle	13.53	29.84
Healthy stroma	15.98	10.96
Fatty tissue	6.74	24.00
Mucus	7.73	5.86
Nerve	0.18	0.40
Stroma lamina propria	7.21	0.55
Healthy glands	6.35	0.75
Background	7.33	2.02

Table 1: Percentage of pixels per class in datasets annotated with sparse and dense annotations.

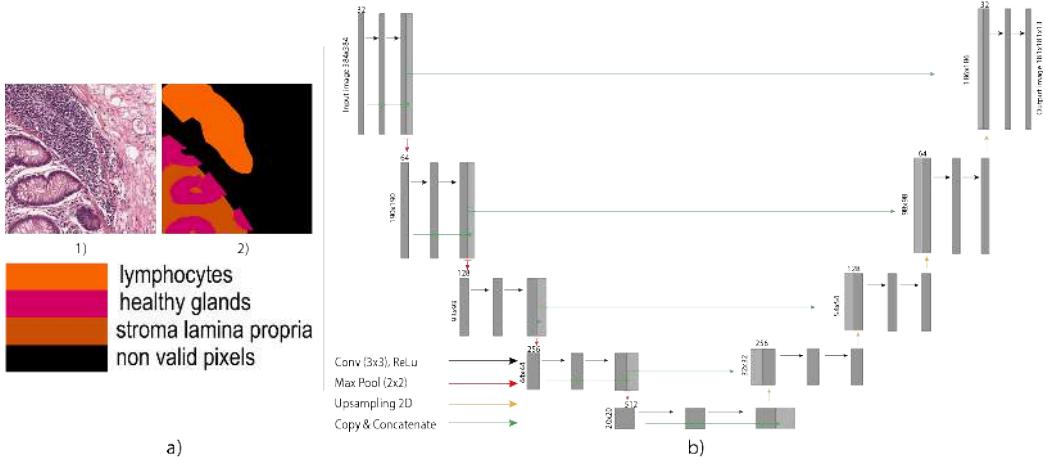


Figure 2: a1) patch from the sparsely annotated dataset with multiple classes, a2) the corresponding annotation map. b) used U-net architecture. Each dark-gray box represents a multi-channel feature map. The input size is shown on the left of the box, and the number of channels on top. The light-gray boxes represent copies of the feature maps. The arrows denote the different operations.

**Mask of valid pixels, no balancing.** We assume that pixels that have not been annotated in the training set should not contribute to the optimization of the network during training. For this purpose, we define a “mask of valid pixels”. In practice, this mask consists of a weight map  $W$  of coefficients  $w_{ij} \in \{0, 1\}$ , where  $w_{ij} = 0$  is used for pixels that are not annotated (label  $y_{ij} = 0$ ), and  $w_{ij} = 1$  is used for pixels that are annotated. In this way, all annotated pixels are considered as “valid” and equally contribute to the loss, not taking into account for a possible class imbalance. We apply this mask to all experiments in this paper, and we also refer to it as a case in which *no balancing* is applied.

**Instance based balancing.** Both in the case of sparse and dense annotations, a single instance can contain multiple classes with a different amount of pixels per class. Let us define as  $L_I$  the amount of valid pixels in an instance (i.e. a training patch) and as  $C_I$  the amount of classes present in that instance. To compensate for class imbalance in a patch, we formulated a weight map that ensures that (1) only valid pixels are considered, and (2) all classes contribute the same to the loss:

$$w_{ij} = \begin{cases} \frac{L_I}{C_I C_{ij}}, & \text{if } y_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where  $C_{ij}$  represents the amount of pixels belonging to the class in position  $(i, j)$ .

**Mini-batch based balancing.** When mini-batch gradient descent is used, an instance based balancing strategy does not take into account the distribution of labels within the mini-batch. In some cases, this may result in some classes having little contribution to parameters update, for example when they only appear in one instance, while other classes may appear in multiple instances of the

	Sparse			Dense			Combined		
	w/o	inst	mb	w/o	inst	mb	w/o	inst	mb
Background	0.32	0.21	<b>0.34</b>	0.25	0.25	0.22	0.24	0.25	0.23
Desmoplastic stroma	<b>0.69</b>	0.68	0.67	<b>0.69</b>	<b>0.69</b>	<b>0.69</b>	0.65	0.63	0.67
Erythrocytes	0.53	0.49	0.62	<b>0.72</b>	<b>0.72</b>	<b>0.72</b>	0.50	0.68	0.65
Fat	0.84	0.77	0.84	0.85	0.84	0.85	0.80	<b>0.86</b>	<b>0.86</b>
Healthy glands	0.83	0.86	0.86	<b>0.88</b>	<b>0.88</b>	0.87	0.84	0.86	0.86
Healthy stroma	0.62	<b>0.67</b>	0.66	0.47	0.48	0.48	0.50	0.64	0.59
Lymphocytes	0.82	<b>0.83</b>	<b>0.83</b>	0.71	0.71	0.72	0.79	0.81	0.76
Mucus	0.22	0.20	0.47	<b>0.96</b>	<b>0.96</b>	<b>0.96</b>	0.42	0.76	0.89
Muscle	0.66	0.61	0.54	0.65	0.65	0.66	<b>0.70</b>	<b>0.70</b>	0.65
Necrosis and Debris	0.28	0.38	0.39	0.41	0.41	0.41	<b>0.42</b>	<b>0.42</b>	0.39
Nerve	<b>0.69</b>	0.64	0.56	0.62	0.61	0.62	0.55	0.62	0.56
Stroma lamina propria	0.76	0.78	0.76	<b>0.82</b>	<b>0.82</b>	<b>0.82</b>	0.77	0.81	0.79
Tumor	<b>0.86</b>	<b>0.86</b>	<b>0.86</b>	<b>0.86</b>	0.85	<b>0.86</b>	0.84	0.85	0.84
Overall	0.61	0.62	0.65	<b>0.68</b>	<b>0.68</b>	<b>0.68</b>	0.62	<b>0.68</b>	0.67

Table 2: Dice scores for every class per annotation type; *w/o* refers to no balancing, *inst* to instance based balancing and *mb* to mini-batch balancing.

same mini-batch. For this reason, we extend the concept of balancing to the mini-batch by defining the amount of valid pixels in a mini-batch as  $L_B$  and the amount of classes in a mini-batch as  $C_B$ . As done for the instance based balancing, each pixel in position  $(i, j)$  contributes to the loss with a coefficient  $w_{ij}$  computed as follows:

$$w_{ij} = \begin{cases} \frac{L_B}{C_B C_{ij}}, & \text{if } y_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where  $C_{ij}$  represents the amount of pixels belonging to the class in position  $(i, j)$ .

**Training.** A five level deep U-Net has been chosen as the segmentation network (see Figure 2b). The network architecture is based on the original U-Net paper (Ronneberger et al., 2015) where the number of filters is doubled after every max-pooling layer and the initial filter size is set to 32. Additionally, skip connections within convolutional layers have been added, where the input of the layer block is concatenated with the last feature map. Transposed convolutions have been replaced with up-sampling operations followed by a convolution in the expansion part.

Multiple U-Net models were trained using sparse annotations, dense annotations and a combination consisting of sparsely annotated images and densely annotated images in a ratio of 4:1. The input of all network configurations was a RGB patch of  $384 \times 384$ px with a pixel size of  $1\mu m$ . For all annotation types all the proposed weight balancing methods were applied.

During training, data was augmented by random flipping, rotation, elastic deformation, blurring, brightness (random gamma), color and contrast changes. An adaptive learning rate scheme was

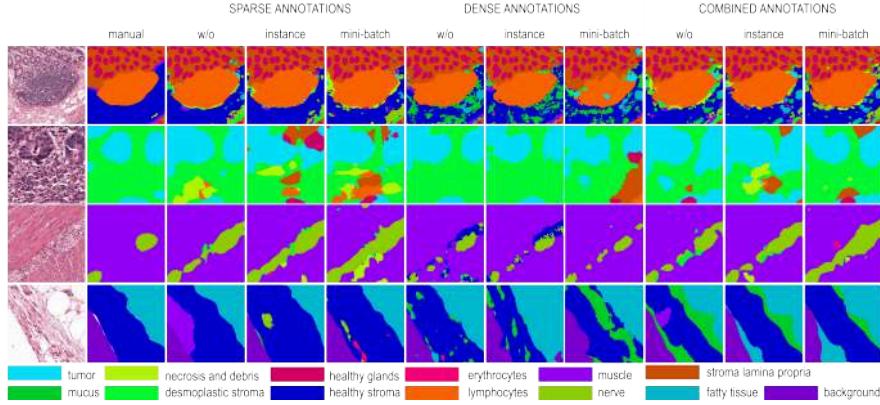


Figure 3: Segmentation output for the considered approaches.

used, where the learning rate was initially set to 0.00001 and then multiplied by a factor of 0.5 after every 10 epoch if no increase in performance was observed on the validation set. The weights of the network were initialized as proposed in (He et al., 2015). The mini-batch size was set to 8 instances per batch, the networks were trained for a maximum of 300 epochs, with 750 iterations per epoch. Categorical cross entropy was used as loss function. The output of all networks is in the form of  $C$  likelihood maps. To obtain a final segmentation output the arg-max was taken as the final label.

## 4. Results

The test set only contained densely annotated regions. From the 5 WSI's used for testing a total of 49 manually annotated regions were selected with a minimum area of  $0.375 \text{ mm}^2$  and a maximum area of  $0.780 \text{ mm}^2$  per region. From these regions 1250 non overlapping tiles were extracted and segmented by the network. The Dice score was used as performance metric. Dice was calculated for every individual class and as a (class) overall score (see Table 2).

Models trained with dense annotations achieved the best performance (Dice = 0.68). The differences across various balancing methods are marginal without a clear preference for any of the balancing methods.

It can be noted that when sparse annotations were used, mini-batch based balancing outperformed instance based balancing slightly. The instance based balance method gives a slight improvement over training without balancing with Dice scores of 0.62 over 0.61 respectively. Applying mini-batch based balancing shows a better added value with a Dice of 0.65.

A similar trend is observed when sparse and dense annotations are combined. In this case, using instance based normalization allows to achieve a Dice = 0.68, which is comparable to what has been obtained with dense annotations. It is worth noting that comparable performance has been achieved with a significantly reduced amount of dense annotations, namely only 20% of dense annotations.

Visual examples of results for the considered approaches and weight balancing strategies are depicted in Figure 3.

## 5. Discussion and conclusions

We have introduced different strategies to modify the loss function in semantic segmentation using a U-Net architecture, in order to address the problem of class imbalance and lack of annotated pixels in training examples, namely (a) instance based balancing and (b) mini batch based balancing. The results show that, in training with sparsely annotated images, only considering valid pixels without introducing any balancing strategies gives the lowest performance. Instance weight balancing slightly improves performance, but this annotation method seems to be best supported by weight balancing at the level of the mini batches. This corroborates the validity of a mini-batch based balancing in cases where for example one single class is only present in an instance, which may be penalized depending on the rest of the instances in the mini-batch.

We experimentally observed that the results for dense annotations are not influenced by any balancing strategy. This can be due to the fact that when all pixels are valid and multiple classes are present in an instance, very little variation is caused to the loss when different strategies are used.

When combined annotations are used, the best result is obtained when instance based balancing is applied. This is in contrast with using only sparse annotations, and can be explained by the fact that balancing at mini-batch level in the presence of a few densely annotated instances in the mini-batch eventually penalizes those annotations, in a pool of multiple sparsely annotated instances with multiple invalid pixels.

If we specifically zoom in on the training scores with the mixed annotated dataset versus the scores on training with the fully annotated set, full dense annotation appears to perform well in the presence of classes that have a clearly visible border with surrounding tissues (as for example tumor or nerve), but predicted maps tend to include multiple classes when segmenting tissues that are more intertwined with neighboring tissue, as can be seen in Figure 3. Objects with a clear boundary (e.g., healthy glands) can be segmented well by most of the approaches.

Based on the proposed balancing methods on the segmentation problem at hand, we can conclude that using sparsely annotated images mixed with a little amount of densely annotated samples allows to get an overall performance that is comparable with using fully annotated instances, in particular when an instance based balancing strategy is applied. More research on different datasets, also in a field different from histopathology, is needed to verify the general validity of the proposed balancing strategies.

## Acknowledgments

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825292, from the Dutch Cancer Society, project number 10602/2016-2, and from the Alpe dHuZes / Dutch Cancer Society Fund, grant number KUN 2014-7032.

## References

Oscar GF Geessink, Alexi Baidoshvili, Joost M Klaase, Babak Ehteshami Bejnordi, Geert JS Litjens, Gabi W van Pelt, Wilma E Mesker, Iris D Nagtegaal, Francesco Ciompi, and Jeroen AWM van der Laak. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. *Cellular Oncology*, pages 1–11, 2019.

- Ben Glocker, Darko Zikic, Ender Konukoglu, David R Haynor, and Antonio Criminisi. Vertebrae localization in pathological spine ct via dense classification from sparse annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 262–270. Springer, 2013.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. Size-constraint loss for weakly supervised cnn segmentation. In *International Conference on Medical Imaging with Deep Learning (MIDL 2018)*, 2018.
- Wilma E Mesker, Jan Junggeburt, Karoly Szuhai, Pieter de Heer, Hans Morreau, Hans J Tanke, and Rob AEM Tollenaar. The carcinoma–stromal ratio of colon carcinoma is an independent factor for survival compared to lymph node status and tumor stage. *Analytical Cellular Pathology*, 29(5):387–398, 2007.
- Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A Rutherford, Joseph V Hajnal, Bernhard Kainz, et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE transactions on medical imaging*, 36(2):674–683, 2017.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical image analysis*, 18(3):591–604, 2014.

# Segmenting Potentially Cancerous Areas in Prostate Biopsies using Semi-Automatically Annotated Data

**Nikolay Burlutskiy**<sup>1</sup>

NIKOLAY.BURLUTSKY@CONTEXTVISION.SE

**Nicolas Pinchaud**<sup>1</sup>

NICOLAS.PINCHAUD@CONTEXTVISION.SE

**Feng Gu**<sup>1</sup>

FENG.GU@CONTEXTVISION.SE

**Daniel Hägg**<sup>1</sup>

DANIEL.HAGG@CONTEXTVISION.SE

**Mats Andersson**<sup>1</sup>

MATS.ANDERSSON@CONTEXTVISION.SE

**Lars Björk**<sup>1,2</sup>

LARS.BJORK@CONTEXTVISION.SE

**Kristian Eurén**<sup>1</sup>

KRISTIAN.EUREN@CONTEXTVISION.SE

**Cristina Svensson**<sup>1</sup>

CRISTINA.SVENSSON@CONTEXTVISION.SE

**Lena Kajland Wilén**<sup>1</sup>

LENA.KW@CONTEXTVISION.SE

**Martin Hedlund**<sup>1</sup>

MARTIN.HEDLUND@CONTEXTVISION.SE

<sup>1</sup> ContextVision AB, Stockholm, Sweden

<sup>2</sup> NORDFERTIL Research Lab Stockholm, Department of Women’s and Children’s Health, Karolinska Institutet and University Hospital

## Abstract

Gleason grading specified in ISUP 2014 is the clinical standard in staging prostate cancer and the most important part of the treatment decision. However, the grading is subjective and suffers from high intra and inter-user variability. To improve the consistency and objectivity in the grading, we introduced glandular tissue WithOut Basal cells (WOB) as the ground truth. The presence of basal cells is the most accepted biomarker for benign glandular tissue and the absence of basal cells is a strong indicator of acinar prostatic adenocarcinoma, the most common form of prostate cancer. Glandular tissue can objectively be assessed as WOB or not WOB by using specific immunostaining for glandular tissue (Cytokeratin 8/18) and for basal cells (Cytokeratin 5/6 + p63). Even more, WOB allowed us to develop a semi-automated data generation pipeline to speed up the tremendously time consuming and expensive process of annotating whole slide images by pathologists. We generated 295 prostatectomy images exhaustively annotated with WOB. Then we used our Deep Learning Framework, which achieved the 2<sup>nd</sup> best reported score in Camelyon17 Challenge, to train networks for segmenting WOB in needle biopsies. Evaluation of the model on 63 needle biopsies showed promising results which were improved further by finetuning the model on 118 biopsies annotated with WOB, achieving F1-score of 0.80 and Precision-Recall AUC of 0.89 at the pixel-level. Then we compared the performance of the model against 17 biopsies annotated independently by 3 pathologists using only H&E staining. The comparison demonstrated that the model performed on a par with the pathologists. Finally, the model detected and accurately outlined existing WOB areas in two biopsies incorrectly annotated as totally WOB-free biopsies by three pathologists and in one biopsy by two pathologists.

## 1. Introduction

The Gleason system was developed in 1966 where architectural patterns are taken as a basis for determining the score ([Gleason and Mellinger, 1974](#)). The system has been revised several times to reflect tumor biology better ([Chen and Zhou, 2016](#)). Currently, ISUP 2014 is the clinical standard in staging prostate cancer and the most important part of the treatment decision. However, the grading suffers from high intra- and inter-user variability, with reported Gleason score discordance ranging from 30-53% ([Nagpal et al., 2018](#)).

The goal of this research is to develop a decision support tool to aid pathologists in their work. We trained Deep Learning (DL) models to segment potentially cancerous areas in prostate biopsies. The models will provide relevant regions of slides for pathologists to focus on, helping them to work more efficiently and save time. Unfortunately, it is difficult to train a DL network on Gleason annotations since it is almost impossible to acquire a consistent Ground Truth for comparison. As a first step in prostate cancer grading, we introduced a class that we objectively have more control of, Glandular tissue WithOut Basal cells (WOB) class. The presence of basal cells is the most accepted biomarker for benign glands and the absence of basal cells is a strong indicator of acinar prostatic adenocarcinoma, the most common form of prostate cancer ([Herawi and I Epstein, 2007](#); [Trpkov et al., 2009](#)). Glandular tissue can objectively be selected as WOB or not WOB by using specific immunostaining for glandular tissue (Cytokeratin 8/18) and for basal cells (Cytokeratin 5/6 + p63) ([Björk et al., 2018](#)). We developed a data generation pipeline to semi automatically produce H&E images with aligned WOB ground truth masks from the scanned immunofluorescence images. One important exception was made for intraductal cancer of the prostate (IDC-P). IDC-P is represented by high grade cancer (Gleason grade 4-5) inside a benign gland with basal cells. IDC-P cases were manually annotated as WOB. Then we used our DL Framework which reached the 2<sup>nd</sup> best reported score in *Camelyon17 Challenge* ([Pinchaud and Hedlund, 2018](#)) to train models on such data.

In this paper, we demonstrate that DL models trained on such data are capable to predict potentially cancerous areas in biopsies with high accuracy. Finally, we compare the performance of the trained model against 17 biopsies annotated independently by three pathologists using only H&E staining. The comparison demonstrate that the model performs as well as the pathologists. Even more, the model detects and accurately outlines existing WOB areas in two biopsies which were incorrectly annotated as totally WOB-free biopsies by three pathologists and in one biopsy by two pathologists.

## 2. Related Work

Recent advances in Artificial Intelligence (AI) and especially in DL have demonstrated high potential for automatic detection and classification of anomalies in Whole Slide Images (WSIs). For example, detection and classification of breast cancer ([Bejnordi et al., 2017](#); [Liu et al., 2017](#); [Pinchaud and Hedlund, 2018](#)), lung cancer ([Coudray et al., 2018](#); [Burlutskiy et al., 2018](#)), colon cancer ([Mohammed et al., 2018](#)), and prostate cancer ([Campanella et al., 2018](#); [Arvaniti et al., 2018](#); [Nagpal et al., 2018](#)) are the tasks where DL systems can help pathologists.

Several AI based systems for predicting anomalies in prostate tissues have been introduced recently. These systems include cancer detection in needle core biopsies ([Campanella et al., 2018](#)), Gleason grading of tissue microarrays ([Arvaniti et al., 2018](#)), and Gleason grading in prostatectomies ([Nagpal et al., 2018](#)).

The authors in (Campanella et al., 2018) used multi-instance learning approach to train a classification network on a large dataset of 12,160 slides of digitalized needle biopsies. The chosen approach allowed to avoid the expensive pixel-wise annotations. However, it is hard to access the performance of the predictions on the pixel-level since the predictions were performed and evaluated on the slide-level. On contrary, in our paper, we produce and evaluate predictions for biopsies on the pixel-level.

The research conducted in (Arvaniti et al., 2018) demonstrated that a DL model trained on Gleason graded tissue microarrays was capable to produce predictions comparable with the inter-pathologist agreement ( $\kappa=0.71$ ). However, the model was trained only on tissue microarrays which are outside of routine clinical workflow of the pathologist; the cores for tissue microarrays represent only small tumor regions and are used for research purposes only. On the other hand, we demonstrate the performance of our DL models on needle biopsies which are of explicit value to the pathologist’s routine clinical workflow.

In (Nagpal et al., 2018) the authors proposed using DL for classifying the Gleason score of WSIs. The diagnostic accuracy of 0.70 ( $p=0.002$ ), higher than for pathologists, was reported. The authors also pointed out that the current Gleason can be refined to more finely characterize and describe tumor morphology quantitatively. The authors used a modified version of *InceptionV3* (Szegedy et al., 2015) to classify input image patches. Also, ensembling and hard example mining were employed to improve performance of trained models. In contrast, we formulate the problem as a binary segmentation task and use unet (Ronneberger et al., 2015) to produce pixel-level predictions. We provide pixels level segmentation results for biopsies while in (Nagpal et al., 2018) the authors concentrated only on prostatectomies and slide-level predictions.

### 3. The Deep Learning Framework

We developed a DL framework using TensorFlow (Abadi et al., 2015) for the training of Deep Neural Networks (DNNs) on Whole Slide Images (WSIs). The framework achieved the 2<sup>nd</sup> best reported score in *Camelyon17 Challenge* (Pinchaud and Hedlund, 2018). The key features of the framework are **quasi online hard example mining**, **compounding of semantic segmentation networks**, and **extensive data augmentation**.

**Quasi Online Hard Example Mining.** A WSI can contain an order of  $10^9$  pixels, making it impossible to train DL networks directly on the full image. Thus, we trained our networks on smaller image patches. It was shown (Nagpal et al., 2018; Pinchaud and Hedlund, 2018) that the patch sampling strategy during training has a great impact on the final performance of the model. We developed a training pipeline that enables *quasi online hard example mining* sampling strategy. We trained our models with patches dynamically extracted from WSIs using a pixel-level probability density function inferred from the error of the model on the pixel-level classification. The probability density function is computed on a separate process synchronously with the training process. This allows the hard example mining to be performed on frequent training cycles efficiently. See Figure A.1 in Appendix for more details.

**Semantic Segmentation Networks.** Semantic segmentation networks predict one of the class labels for every pixel of a patch. While our framework can handle a large variety of segmentation networks, in this study we used the state-of-the-art **unet** (Ronneberger et al., 2015). This model has an encoder for downsampling and a decoder for upsampling, which are linked with lateral

skip connections. The network has high capacity of feature representation, thus, it can learn and aggregate knowledge at multiple scales in the data. Unet has become one of the most popular networks for semantic segmentation problems in biomedical imaging ([Burlutskiy et al., 2018](#); [Li et al., 2018](#); [Pinchaud and Hedlund, 2018](#)).

**Using Compound Model.** In order to increase the receptive field of the network, we developed a compound model approach which allows to expose the trained model to different resolution levels in WSIs. We combine two **unet** models learned from 1 and 2 mpp<sup>1</sup> resolutions. Let  $M^1$  be a **unet** model trained on 1 mpp. The network maps an RGB pixel  $i \in [0, 255]^3$  to the probability  $M^1(i) \in [0, 1]$  of belonging to the WOB class. Then let  $M^{1,2}$  be another **unet** model trained on a resolution of 2 mpp using the output of  $M^1$  as an extra input channel as illustrated in Figure [A.3](#). Compound models described above bring the benefits of ensemble learning: we can learn the individual **unet** models with different strategies or hyper-parameters inducing different expressiveness of models that can eventually be integrated as their combination in the compound model.

**Data Augmentation.** The difficulty of producing large amount of annotated data can be leveraged with an extensive usage of data augmentation. Our framework allows to program a data processing pipeline in a simple text based configuration file. Different transformation operators can be piped to produce a final augmented patch. The augmentation is performed online during the training. The augmentation operators include rotation, mirroring, elastic deformation, color jittering. Table [A.1](#) in Appendix provides augmentation examples.

## 4. Data Generation

In order to obtain a more objective ground truth, we introduced the concept of WOB, ‘glandular tissue WithOut Basal cells’. The objectivity of WOB is in the fact that the presence of basal cells can be assessed by using immunohistochemical markers. Presence of basal cells indicates that a gland is healthy which implies that WOB corresponds to potentially malignant or cancerous structures ([Herawi and I Epstein, 2007](#); [Trpkov et al., 2009](#)). For our experiments, we produced three WOB annotated datasets described below (more details on the datasets are in Table [A.2](#)).

### 4.1. Semi-automatically WOB annotated prostatectomies

WOB areas can be accurately segmented out in the prostate tissues using a method described in ([Euren, 2016](#); [Björk et al., 2018](#)). The method is based on staining prostate tissues towards basal cells with Cytokeratin 5/6, then epithelial tissue with Cytokeratin 8/18, cancerous cells with Alpha-methylacyl-CoA racemase (AMACR), and nuclei with 4’,6-diamidino-2-phenylindole (DAPI) ([Björk et al., 2018](#)). Then the same slides are stained with H&E and scanned at a high resolution. The immunofluorescent stainings mark specific structures in the prostate tissues which can be automatically converted into the WOB areas. By overlaying the stainings with H&E, the WOB areas exactly match the corresponding WOB areas in an H&E image which allows us to generate accurate and detailed WOB annotations.

**Converting immunofluorescence images into annotations.** We developed an algorithm to generate binary ground truth masks from scanned immunofluorescence images. For the different immunofluorescence channels, the information are not present in the entire gland areas but localized

---

1. micrometer per pixel

to certain positions within the gland. As a result, the markers of the different immunochannels consequently appear at different places within the gland. To make an overall segmentation of the entire gland area, this local information is propagated over the gland by the density estimation filter which allowed to distribute the local information evenly within a local neighborhood of the image. The density estimation filter was applied to all three immunofluorescence channels (Cytokeratin 5/6, Cytokeratin 8/18, and AMACR). Then two heatmaps were generated. The first one mapped epithelial versus epithelial plus basal channel. The second mapped AMACR versus epithelial plus AMACR channel. Both heatmaps produced an estimate in the range [0, 1] where high intensity indicated high probability for the WOB class. The design of the heatmaps is based on the fact that adenocarcinoma only occurs within epithelial tissue and the relation between the epithelial channel to the basal cell and AMACR channels consequently indicate the WOB or not WOB class. Finally, these two heatmaps were combined to generate a best possible WOB mask (see Figure A.2 for a WOB mask example). To our experience the AMACR heatmap was less reliable compared to the basal cell heatmap and when they were inconsistent the basal cell heatmap was used.

**Reviewing the automatically generated annotations.** We decided to ask pathologists to review the automatically generated masks. The main reason was that the intraductal cancer of the prostate (IDC-P) represented by high grade cancer (Gleason grade 4-5) inside a benign gland with basal cells must be manually annotated. The reviews led to some minimal manual corrections in the automatically generated masks. In total, 48 prostatectomy masks were generated and reviewed.

**Using consecutive slices.** We included 48 consecutive H&E stainings to the original 48 H&E prostatectomies in order to further increase the size of trained data and eventually the performance of trained models. A few consecutive slices were excluded since the difference in morphology was not acceptably negligible. The ground truth for the consecutive images was obtained by registering the corresponding H&E image and then applying the calculated transformation to the ground truth of the original H&E images (see Figure A.4 for an example). The registration was done using non rigid registration with elastix software (<http://elastix.isi.uu.nl/>).

**Extending scanners variation.** All H&E stainings were scanned with three commercially available scanners. The reason for scanning the images by the three scanners was the fact that we wanted the model to be robust across these scanners. The scanned images were registered to the original images and a corresponding ground truths were produced.

**Final dataset.** We produced 295 images exhaustively annotated with WOB; the images represented variations in prostate cancer stages, scanners, and tissue morphology. This dataset was used for training DL models.

#### 4.2. Manually annotated WOB biopsies supported by H&E and DAB stainings

The second dataset consisted of 181 biopsies scanned with two different scanners at 0.5 and 0.22 micrometer per pixel accordingly. All biopsies were exhaustively annotated by six pathologists. All the pathologists who annotated the images were carefully selected. The criteria was that the pathologists had proper specialist training in pathology as well as several years of clinical practice in pathology. The number of years of experience varied between 5-20 for both groups. The pathologists used AMACR, a staining for cancerous cells, and p63, a staining for basal cells. This dataset was split into train set of 118 biopsies and 63 biopsies for test. The train set was used for training and finetuning DL models and the test set was used for evaluation of the trained models.

### 4.3. Manually annotated WOB biopsies supported by H&E staining only

Finally, 17 out of 181 annotated biopsies, 13 with WOB and 4 without WOB, were exhaustively annotated by other three pathologists with H&E staining only, without any support of p63 and AMACR stainings. This dataset was used for comparison of DL models to three pathologists.

## 5. Experimental Setup and Results

In total, we trained and evaluated six DL models on a cluster with 10 TitanXp GPUs with 12GB VRAM and 64GB RAM. Training all the six models required approximately five days.

### 5.1. Trained DL models

We trained DL models (1) only on prostatectomies, (2) only on biopsies, and (3) on prostatectomies first and then finetuned on biopsies. For each dataset we trained two models, a model  $M^1$  on 1 mpp and a *compound model*  $M^{1,2}$  on 1 mpp and 2 mpp (see Figure A.3).

The model  $M_{pr}^1$  was trained only on 1 mpp prostatectomies,  $M_{pr}^{1,2}$  was trained on 1 and then 2 mpp prostatectomies; the model  $M_{bi}^1$  and  $M_{bi}^{1,2}$  were trained only on biopsies on 1 mpp and on both 1 mpp and 2 mpp correspondingly. Finally,  $M_{pr,bi}^1$  and  $M_{pr,bi}^{1,2}$  were trained on prostatectomies first and then finetuned on biopsies at 1 mpp and then both 1 and 2 mpp.

Each model was trained for  $10^6$  iterations. Each iteration had a batch of 32 patches. The patches were sampled using *quasi online hard example mining* described in Appendix A.1. Each patch was 188 by 188 pixels. The relatively small size of the patches constrained the sampling areas of hard regions without overshooting these.

### 5.2. Evaluation setup

The test set for the evaluation consisted of 63 needle biopsies annotated with WOB by several pathologists with support of DAB stainings. We calculated precision-recall curves, areas under the precision-recall curve (PR AUC), and maximum F1 scores for each model. The results are in Figure 1.

Then we chose the best performing DL model, the compound model  $M_{pr,bi}^{1,2}$ , and compared its performance to three pathologists who manually annotated WOB areas in 17 biopsies using H&E stainings only. The results are summarized in Figure 2.

### 5.3. Performance of DL models

The predictions demonstrated high performance at pixel-level reaching F1 score of 0.80 and PR AUC of 0.89 for the model  $M_{pr,bi}^{1,2}$  trained on prostatectomies and then finetuned on biopsies (see Figure A.5 for prediction examples by different models). We noticed that training compound models on 2 mpp consequently after training models on 1 mpp helped to remove false positives and to improve the performance of the models. The performance of the models  $M_{pr}^1$  and  $M_{pr}^{1,2}$  trained only on prostatectomies showed the worst results with PR AUCs of 0.62 and 0.68 accordingly (see Figure 1, the left plot). The models  $M_{bi}^1$  and  $M_{bi}^{1,2}$  trained only on biopsies achieved higher PR AUCs of 0.71 and 0.78. Finally, the models  $M_{pr,bi}^1$  and  $M_{pr,bi}^{1,2}$  outperformed both biopsies and prostatectomies with PR AUCs of 0.83 and 0.89. The same order holds for F1 scores across the models. Also, the models trained on prostatectomies first and then finetuned on biopsies showed the widest F1 scores

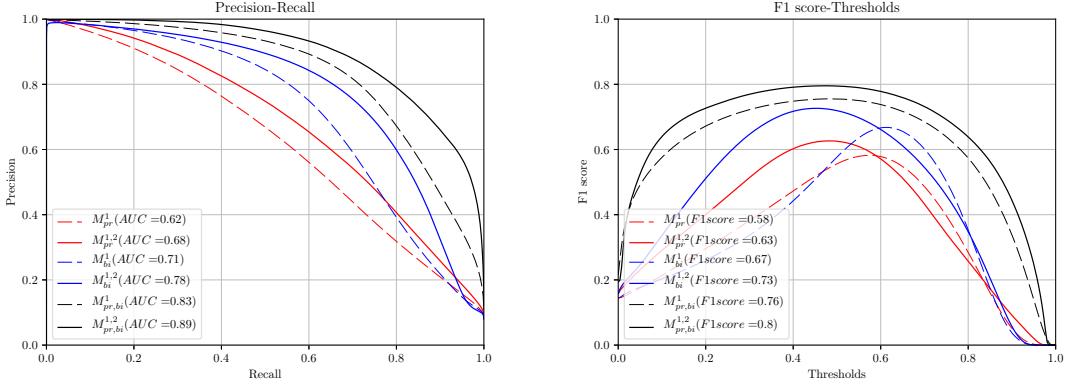


Figure 1: *Precision-Recall (PR) curves* (left) and *F1 score curves* (right) for all models. In the left plot, AUC stands for Area Under PR Curve (AUC) and is calculated for each model. In the right plot, a maximum achievable F1 score for each model is shown.  $M_{pr}^1$  was trained only on 1 mpp prostatectomies,  $M_{pr}^{1,2}$  was a compound model trained on 1 and then 2 mpp prostatectomies; the model  $M_{bi}^1$  and  $M_{bi}^{1,2}$  were trained only on biopsies on 1 mpp and on both 1 mpp and 2 mpp correspondingly. Finally,  $M_{pr,bi}^1$  and  $M_{pr,bi}^{1,2}$  were trained on prostatectomies first and then finetuned on biopsies at 1 mpp and then both 1 and 2 mpp.

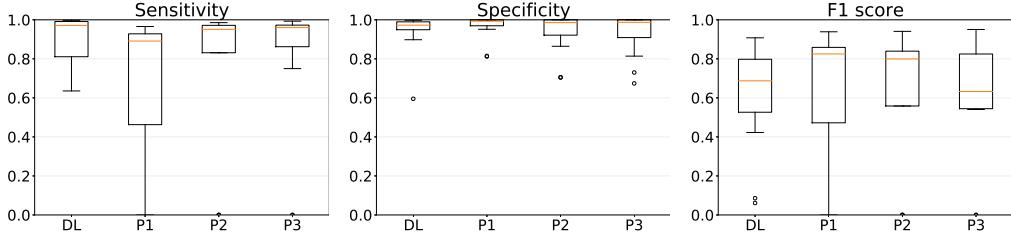


Figure 2: *Sensitivity, specificity, and F1 score* for WOB predicted by the model  $M_{pr,bi}^{1,2}$  (DL) and annotated by the three pathologists (P1, P2, P3) for 17 biopsies. Sensitivity and F1 score were calculated only for 13 WOB positive biopsies while specificity for all 17 biopsies.

(see Figure 1, the right plot) which implies that the models  $M_{pr,bi}^1$  and  $M_{pr,bi}^{1,2}$  are more robust to different thresholds.

#### 5.4. Comparison of the best performing DL model to three pathologists

Finally, we compared the predictions by the best performing model, the compound model  $M_{pr,bi}^{1,2}$ , with three pathologists who were instructed about the WOB class and then annotated WOB areas using only H&E stainings. Remarkably, in two biopsies all three pathologists totally missed WOB areas but our model successfully detected those areas (see Figure A.6 in Appendix for examples of the missed areas). Two out of three pathologists missed all WOB areas in a third biopsy. Quantita-

tively, pathologists performed as well as the trained model (see Figure 2 for sensitivity, specificity, and F1 score boxplots).

## 6. Conclusions and Future Work

In this paper, we introduced a novel DL framework for segmenting potentially cancerous areas in prostate biopsies using semi-automatically generated data. Introducing WOB class allowed us to generate data with minimal involvement of pathologists for pixel-level annotations.

The performance of the trained models was tested on 63 biopsies from several clinical labs, scanners, and annotated by different pathologists. The trained models showed high potential in identifying potentially cancerous areas. We demonstrated that training the proposed compound models is beneficial for increasing the receptive field of the networks and eventually for the performance of the models. We showed that using semi-automatically generated data leads to increase in performance compared to models, trained only on biopsies.

We evaluated the best performing model against three pathologists who annotated 17 biopsies using H&E staining only. The comparison demonstrated that the model performed on a par with the pathologists. Even more, the model detected and outlined existing WOB areas in three biopsies which were annotated as WOB-free biopsies by the three pathologists.

For future work, we plan to evaluate the performance of the trained models at a few clinical pathology labs. We are planning to collect feedback from pathologists who will use our product with the developed DL algorithm. Also, we are going to test the trained models on more diverse WSIs images stained and scanned in different labs in order to evaluate the robustness of the algorithm further. Finally, we are planning to build a mapping from WOB predictions to Gleason score grading scheme.

## Authors' Contributions

N.P. developed the DL framework. N.B. and N.P. designed, carried out the experiments, and wrote the manuscript; N.B. evaluated the results. F.G. helped with annotation generation. M.A. developed the method to generate the annotations. M.A. and D.H. contributed to data augmentation and experiment design. L.B. and K.E. introduced WOB concept, designed and organized data generation process. M.H., L.K., and C.S. supervised the project.

## Acknowledgments

We would like to thank pathologists for providing the data and expertise to the project. We thank David Buffoni for the comments on the manuscript.

## References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah,

Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. URL <http://tensorflow.org/>. Software available from tensorflow.org.

Eirini Arvaniti, Kim Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter Wild, Jan Rüschoff, and Manfred Claassen. Automated Gleason Grading of Prostate Cancer Tissue Microarrays via Deep Learning. *Scientific Reports*, 8(1):12054, 2018. ISSN 2045-2322. URL <https://doi.org/10.1038/s41598-018-30535-1>.

Ehteshami Bejnordi, Veta M, Johannes van Diest P, and et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer. *The Journal of the American Medical Association (JAMA)*, 318(22):2199–2210, 2017. URL <http://dx.doi.org/10.1001/jama.2017.14585>.

Lars Björk, Jonas Gustafsson, Feria Hikmet Noraddin, Kristian Eurén, and Cecilia Lindskog. A New High-Throughput Auto-Annotation Method to Detect and Outline Cancer Areas in Prostate Biopsies. In *14th European Congress on Digital Pathology and the 5th Nordic Symposium on Digital Pathology*, 2018. URL <http://www.ecdp2018.org/#9WOuOWYoLbFKcwuG8n>.

Nikolay Burlutskiy, Feng Gu, Lena Kajland Wilen, Max Backman, and Patrick Micke. A Deep Learning Framework for Automatic Diagnosis in Lung Cancer. *CoRR*, abs/1807.10466, 2018. URL <http://arxiv.org/abs/1807.10466>.

Gabriele Campanella, Vitor Werneck Krauss Silva, and Thomas J. Fuchs. Terabyte-Scale Deep Multiple Instance Learning for Classification and Localization in Pathology. *CoRR*, abs/1805.06983, 2018. URL <http://arxiv.org/abs/1805.06983>.

Ni Chen and Qiao Zhou. The Evolving Gleason Grading System. *Chinese Journal of Cancer Research*, 28(1):58–64, Feb 2016. ISSN 1000-9604. URL <https://www.ncbi.nlm.nih.gov/pubmed/27041927>.

Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images using Deep Learning. *Nature Medicine*, 24(10):1559–1567, 2018. URL <https://doi.org/10.1038/s41591-018-0177-5>.

Kristian Euren. Method and System for Detecting Pathological Anomalies in a Digital Pathology Image and Method for Annotating a Tissue Slide, 06 2016. URL <https://patents.google.com/patent/US20170372471A1/en>. US 15195526.

DF Gleason and GT Mellinger. Prediction of Prognosis for Prostatic Adenocarcinoma by Combined Histological Grading and Clinical Staging. *Journal of Urology*, 111:58–64, 1974. URL <https://www.ncbi.nlm.nih.gov/pubmed/4813554>.

Mehsati Herawi and Jonathan I Epstein. Immunohistochemical Antibody Cocktail Staining (p63/HMWCK/AMACR) of Ductal Adenocarcinoma and Gleason Pattern 4 Cribriform and Non-cribriform Acinar Adenocarcinomas of the Prostate. *The American Journal of Surgical Pathology*, 31:889–94, 07 2007. URL <https://www.ncbi.nlm.nih.gov/pubmed/17527076>.

Jiayun Li, Karthik V. Sarma, King Chung Ho, Arkadiusz Gertych, Beatrice S. Knudsen, and Corey W. Arnold. A Multi-Scale U-Net for Semantic Segmentation of Histological Images from Radical Prostatectomies. *Proceedings of AMIA Annual Symposium*, 2017:1140–1148, Apr 2018. ISSN 1942-597X. URL <https://www.ncbi.nlm.nih.gov/pubmed/29854182>.

Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Nelson, Gregory Corrado, Jason Hipp, Lily Peng, and Martin Stumpe. Detecting Cancer Metastases on Gigapixel Pathology Images. *CoRR*, abs/1703.02442, 2017. URL <http://arxiv.org/abs/1703.02442>.

Ahmed Kedir Mohammed, Sule Yildirim, Ivar Farup, Marius Pedersen, and Øistein Hovde. Y-Net: A Deep Convolutional Neural Network for Polyp Detection. *CoRR*, abs/1806.01907, 2018. URL <http://arxiv.org/abs/1806.01907>.

Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L. Smith, Arash Mohtashamian, James H. Wren, Gregory S. Corrado, Robert MacDonald, Lily Peng, Mahul B. Amin, Andrew J. Evans, Ankur R. Sangoi, Craig H. Mermel, Jason D. Hipp, and Martin C. Stumpe. Development and Validation of a Deep Learning Algorithm for Improving Gleason Scoring of Prostate Cancer. *CoRR*, abs/1811.06497, 2018. URL <http://arxiv.org/abs/1811.06497>.

Nicolas Pinchaud and Martin Hedlund. Camelyon17 Grand Challenge. *Technical Report*, pages 1–3, 05 2018. URL <https://camelyon17.grand-challenge.org/evaluation/results>.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR*, abs/1505.04597, 2015. URL <http://arxiv.org/abs/1505.04597>.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *CoRR*, abs/1512.00567, 2015. URL <http://arxiv.org/abs/1512.00567>.

Kiril Trpkov, Joanna Bartczak-McKay, and Asli Yilmaz. Usefulness of Cytokeratin 5/6 and AMACR Applied as Double Sequential Immunostains for Diagnostic Assessment of Problematic Prostate Specimens. *American Journal of Clinical Pathology*, 132(2):211–220, 2009. URL <https://www.ncbi.nlm.nih.gov/pubmed/19605815>.

## Appendix A. Supplementary Material

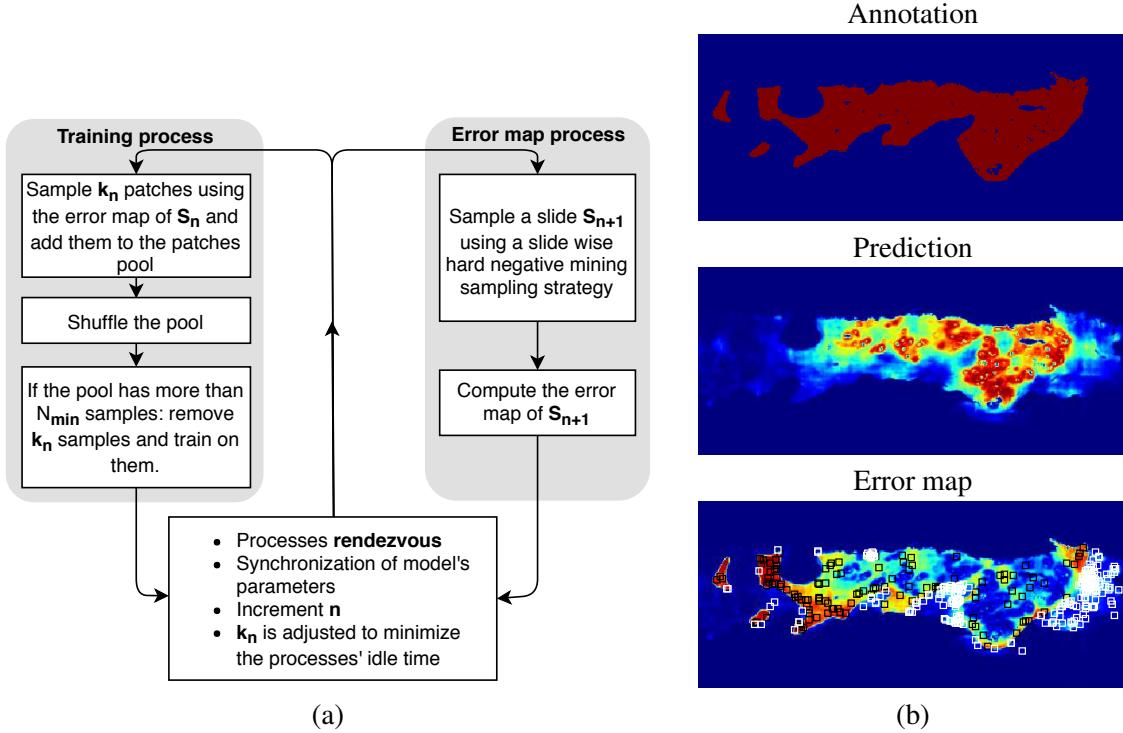
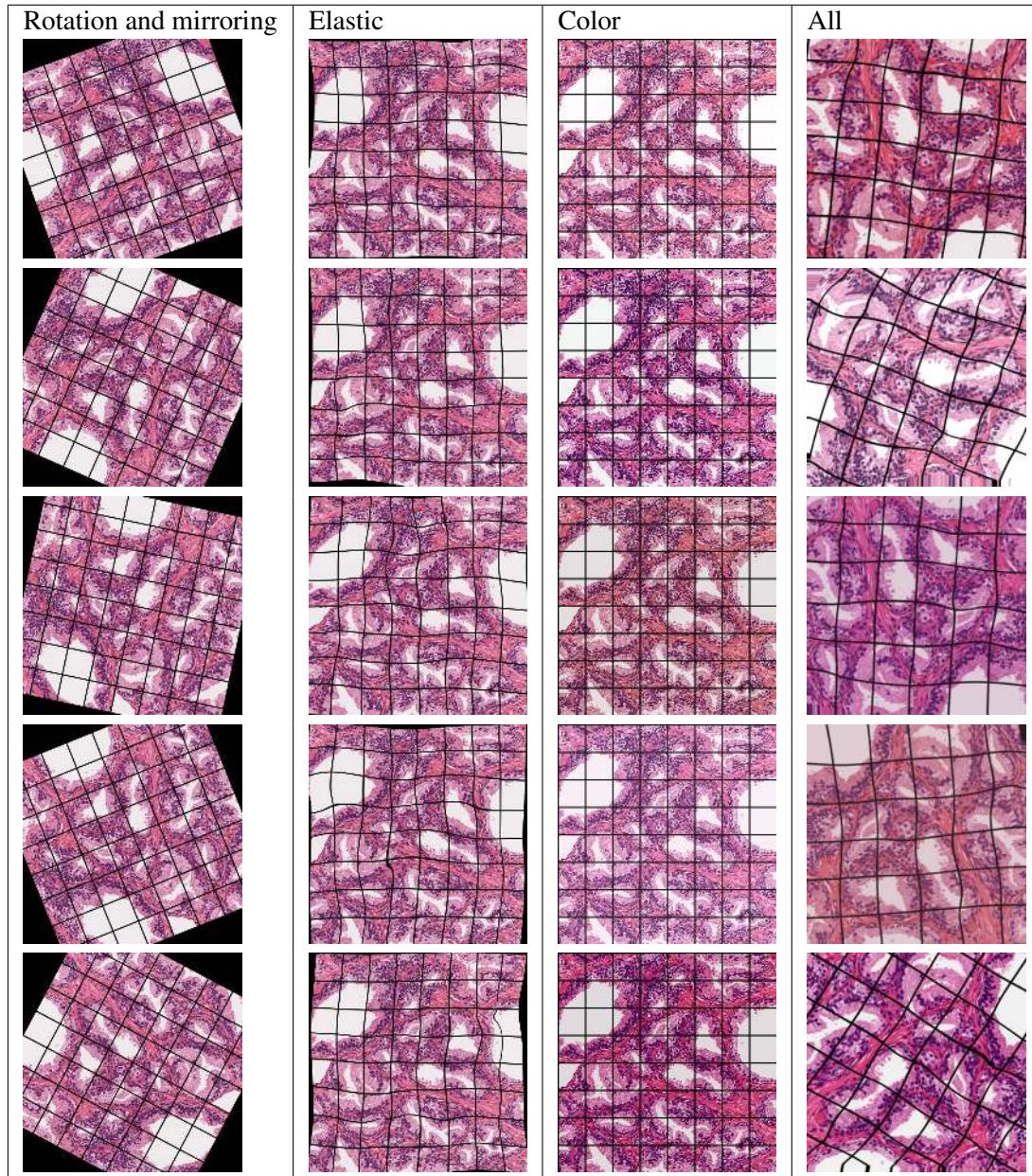


Figure A.1: (a) A flowchart explaining the training with quasi online hard example mining of patches. The **error map process** starts by computing the error map of an initial  $S_0$  slide. Then the **training process** starts and uses the computed error map to sample  $k_0$  patches from  $S_0$  that are fed into a pool. The two processes rendezvous and the cycle continues for  $S_1, S_2, \dots$  until the pool reaches a minimal size  $N_{min}$  to allow the building of mini-batches containing uncorrelated patches. Then the **training process** starts to update the model parameters by training on  $k_n$  patches sampled from the pool. Once the **training process** meets the **error map process** at the rendezvous, the model parameters are synchronized. Then a new cycle begins, the **error map process** samples  $S_{n+1}$  and computes its error map with the updated parameters while the **training process** samples patches from  $S_n$ . The **error map process** samples  $S_{n+1}$  using a slide wise hard example mining strategy, e.g. a slide among those containing many false positives is sampled. At each rendezvous,  $k_n$  is the adjusted version of  $k_{n-1}$  to minimize the idle time. Depending on which process is idling,  $k_n$  is increased for the **training process** or decreased for the **error map process**. The two processes can run on different hardware devices and only need to communicate over the network. (b) An example of a corresponding annotation mask, prediction heat map and error map for a biopsy area. The error map is overlaid with examples of randomly sampled patches for *not WOB* (white) and *WOB* (black) classes. The patches are more likely to be sampled from high error regions.

Table A.1: Augmentations applied on a patch. An overlaid grid illustrates the deformations.



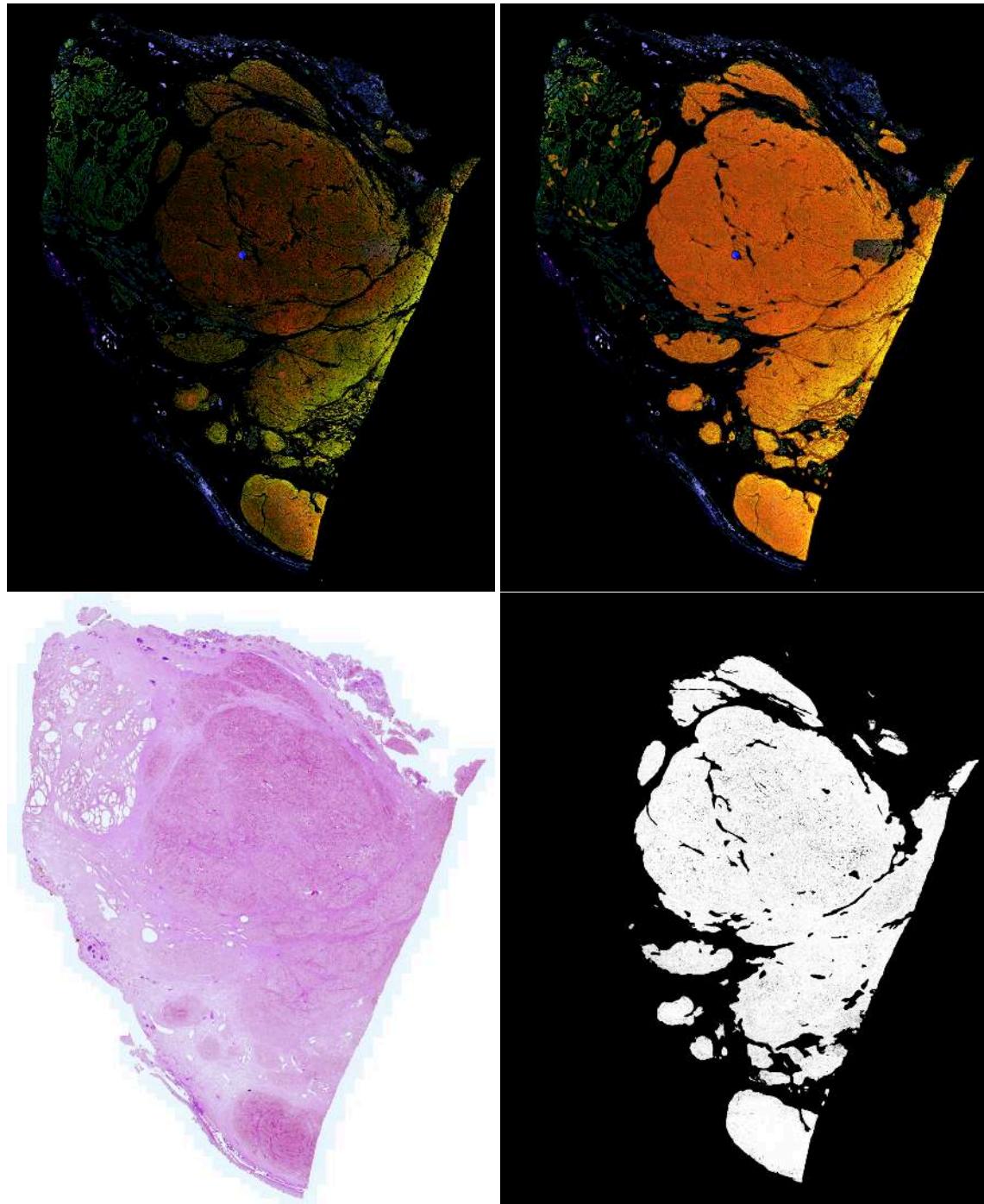


Figure A.2: An example of immunofluorescence channels: red - AMACR, green - epithelial channel, blue - basal cells (top left), a WOB heatmap (orange) generated from the channels overlaid on the immunofluorescence channels (top right), a corresponding H&E staining (bottom left), and a final binary WOB mask (bottom right) for a prostatectomy.

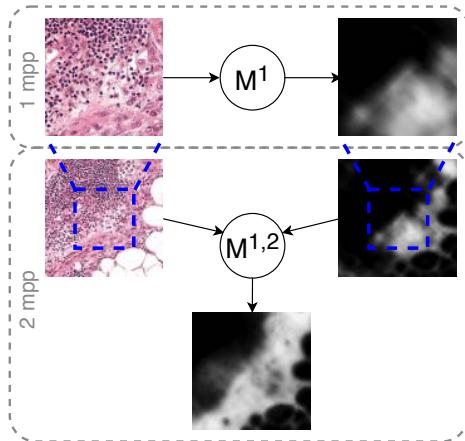


Figure A.3: An illustration of our compound model. A first **unet**  $M^1$  is trained on 1 mpp. A second **unet**  $M^{1,2}$  is trained on 2 mpp by taking the output of  $M^1$  as an extra channel.

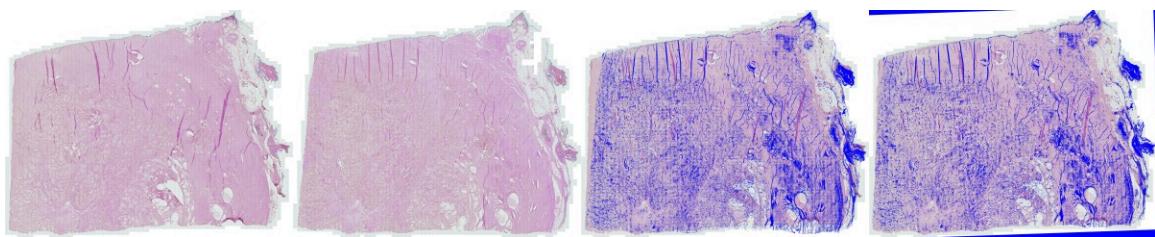


Figure A.4: Examples of registration for original and consecutive prostatectomies. The first image is the original image and the second image is the consecutive slice. The third and the fourth images are the original image overlaid with an outline of the consecutive before and after registration.

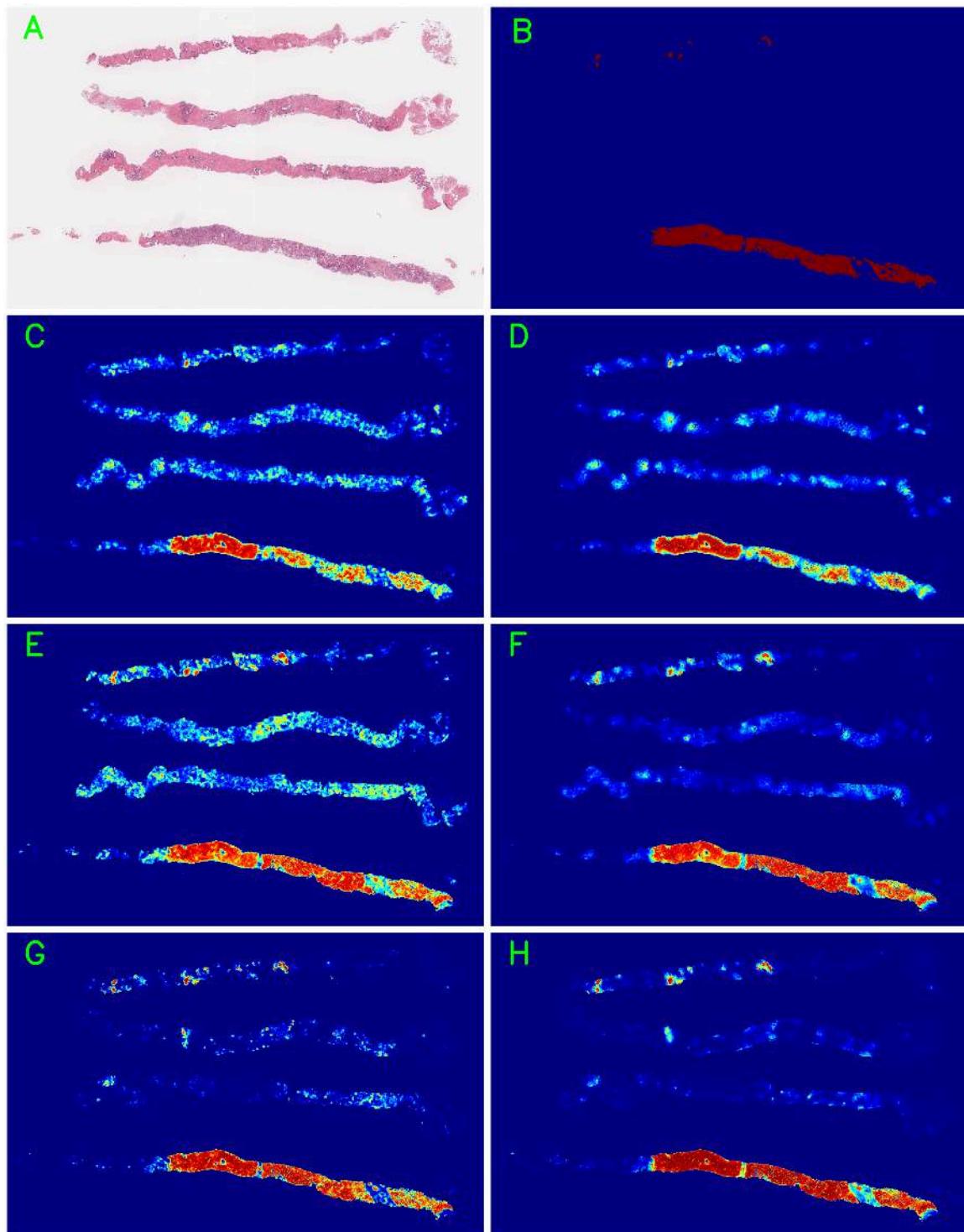


Figure A.5: An example of predictions and the ground truth for a biopsy. (A) is an H&E staining, (B) is an annotation by a pathologist, then there are predictions by (C)  $M_{pr}^1$ , (D)  $M_{pr}^{1,2}$ , (E)  $M_{bi}^1$ , (F)  $M_{bi}^{1,2}$ , (G)  $M_{pr,bi}^1$ , and (H)  $M_{pr,bi}^{1,2}$ .

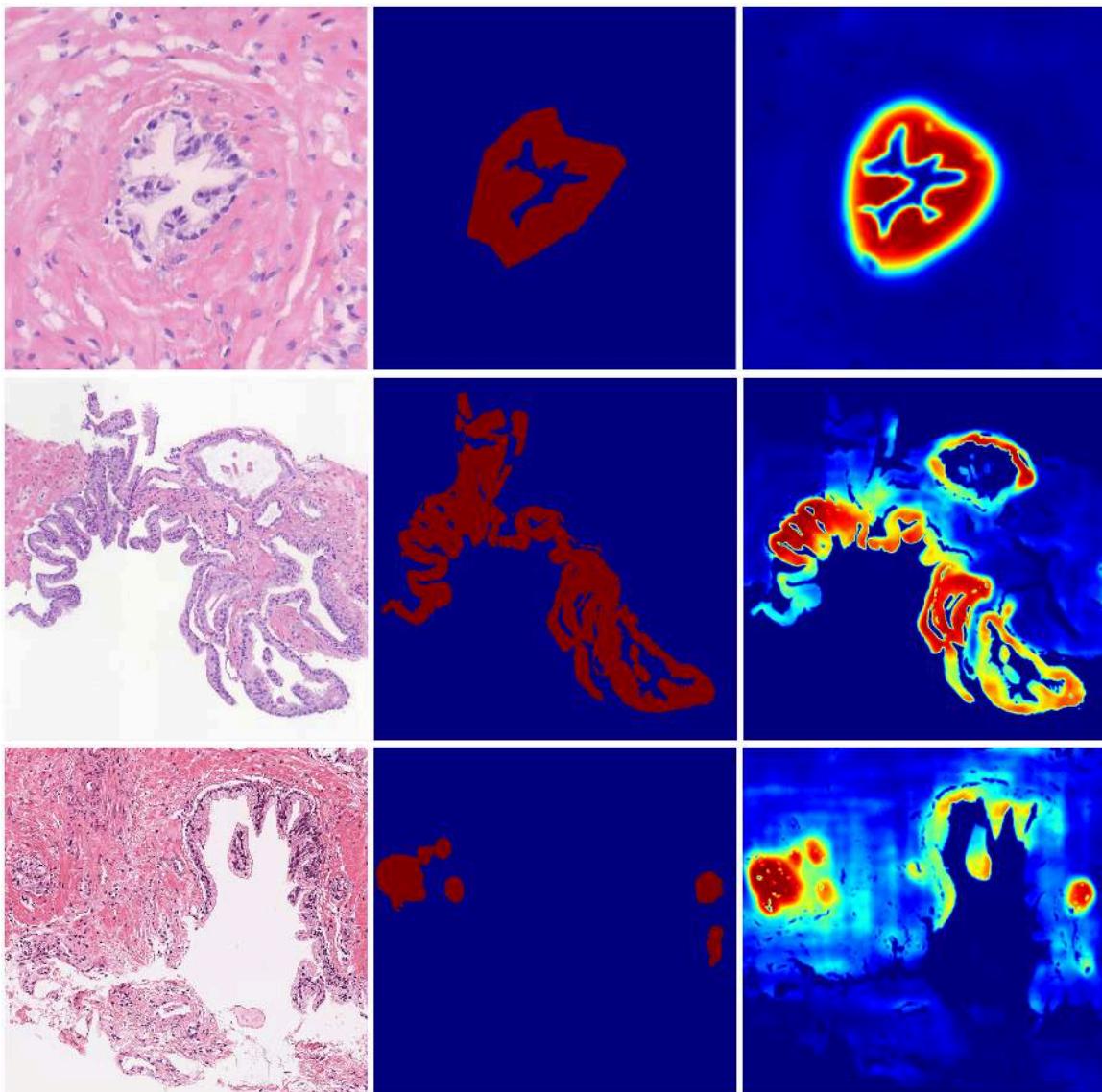


Figure A.6: Examples of missed by three pathologists WOB areas from three biopsies (arranged row-wise). In each row, there is an H&E area first, then a corresponding WOB annotation, and finally a prediction by  $M^{1,2}_{pr,bi}$ . For the first biopsy, two out of three pathologists totally missed all WOB areas, and one pathologist found WOB areas with Sensitivity of 0.75 and Specificity of 1.00. For the second and the third biopsies, all the three pathologists missed all WOB areas in the biopsies. On contrary, our models detected and accurately outlined most of them.

Table A.2: Information on datasets used for training and testing. Prostatectomies were scanned with three different scanners A, B, C and biopsies with D, E. \*consecutive slices of tissues; \*\*Grade Group 0 is for benign samples.

#	Type	Scanner	Train						Test							
			$N_{im}$	Grade Group**						$N_{im}$	Grade Group**					
				5	4	3	2	1	0		5	4	3	2	1	0
#1	prostatect.	A	40	8	1	12	29	-	-	10	2	-	1	7	-	-
#2	prostatect.	A*	43	10	1	12	33	-	-	13	4	-	1	8	-	-
#3	prostatect.	B	40	8	1	12	29	-	-	10	2	-	1	7	-	-
#4	prostatect.	B*	41	11	1	11	32	-	-	14	4	-	1	9	-	-
#5	prostatect.	C	35	9	1	10	24	-	-	9	2	-	1	6	-	-
#6	prostatect.	C*	30	7	1	9	23	-	-	10	2	-	1	7	-	-
Total			229	53	6	66	170	-	-	66	16	-	6	44	-	-
#7	biopsies	D	21	20						1	7	7				-
#8	biopsies	E	97	81						16	56	38				18
Total			118	101						17	63	45				18

# Deep Hierarchical Multi-label Classification of Chest X-ray Images

**Haomin Chen**<sup>1,2</sup>

HCHEN135@JHU.EDU

**Shun Miao**<sup>1</sup>

SHWINMIAO@GMAIL.COM

**Daguang Xu**<sup>1</sup>

DAGUANGX@NVIDIA.COM

**Gregory D. Hager**<sup>2</sup>

HAGER@CS.JHU.EDU

**Adam P. Harrison**<sup>1</sup>

ADAM.P.HARRISON@GMAIL.COM

<sup>1</sup> NVIDIA AI-*Infra*, Bethesda, MD

<sup>2</sup> Department of Computer Science, Johns Hopkins University, Baltimore, MD

## Abstract

Chest X-rays (CXR) are a crucial and extraordinarily common diagnostic tool, leading to heavy research for Computer-Aided Diagnosis (CAD) solutions. However, both high classification accuracy and meaningful model predictions that respect and incorporate clinical taxonomies are crucial for CAD usability. To this end, we present a deep Hierarchical Multi-Label Classification (HMLC) approach for CXR CAD. Different than other hierarchical systems, we show that first training the network to model conditional probability directly and then refining it with unconditional probabilities is key in boosting performance. In addition, we also formulate a numerically stable cross-entropy loss function for unconditional probabilities that provides concrete performance improvements. To the best of our knowledge, we are the first to apply HMLC to medical imaging CAD. We extensively evaluate our approach on detecting 14 abnormality labels from the PLCO dataset, which comprises 198,000 manually annotated CXRs. We report a mean Area Under the Curve (AUC) of 0.887, the highest yet reported for this dataset. These performance improvements, combined with the inherent usefulness of taxonomic predictions, indicate that our approach represents a useful step forward for CXR CAD.

**Keywords:** hierarchical multi-label classification, chest x-ray, computer aided diagnosis.

## 1. Introduction

Chest X-rays (CXR) are the most frequently ordered image study (Folio, 2012). Commensurate with this importance, CXR Computer-Aided Diagnosis (CAD) has received considerable research attention, both prior to the popularity of deep learning (Jaeger et al., 2013), and afterwards (Wang et al., 2017; Yao et al., 2017; Guendel et al., 2018). These efforts have achieved notable successes, e.g., Guendel et al. (2018) reporting very high mean Area Under the Curves (AUCs) on the Prostate, Lung, Colorectal and Ovarian (PLCO) dataset (Gohagan et al., 2000). Yet, pushing raw performance further will likely require models that depart from standard multi-label classifiers. Perhaps more importantly, standard multi-label classifiers are not able to leverage or align with domain knowledge. For instance, despite their importance to clinical understanding and interpretation, taxonomies of disease patterns are not typically incorporated into CXR CAD systems, or for other medical CAD domains for that matter. This observation motivates our work, which uses Hierarchical Multi-Label Classification (HMLC) to both push raw AUC performance further and also to provide more meaningful predictions that leverage clinical taxonomies.

Organizing diagnoses or observations into ontologies and/or taxonomies is crucial within radiology, *e.g.*, RadLex ([Langlotz, 2006](#)), with CXR interpretation being no exception ([Folio, 2012](#); [Demner-Fushman et al., 2015](#); [Dimitrovski et al., 2011](#)). This importance should also be reflected within CAD systems. For instance, when uncertain about fine-level predictions, *e.g.*, *nodules* vs. *masses*, a CAD system should still be able to provide meaningful parent-level predictions, *e.g.*, *pulmonary nodules and masses*. This parent prediction may be all the clinician is interested in anyway. Another important benefit is that observations are conditioned upon their parent being true, allowing fine-level predictors to focus solely on discriminating between siblings rather than on having to discriminate across all possible conditions. This can help improve classification performance ([Bi and Kwok, 2015](#)).

Because more than one abnormality can be observed on a CXR at the same time, a CAD system must operate in a multi-label setting. Prior work has well articulated the limitations of Binary Relevance (BR) learning ([Dembczyński et al., 2012](#)), *i.e.*, treating each label as an independent prediction. HMLC helps address this, by making predictions conditionally independent rather than globally independent. Inferring risk-optimal binary HMLC labels given a set of predictions is a surprisingly rich topic ([Bi and Kwok, 2015](#)), but here we focus instead on producing said predictions. In this way, our focus has similarities to recent deep neural network approaches for hierarchical *multi-class* classification of natural images ([Redmon and Farhadi, 2017](#); [Roy et al., 2018](#); [Yan et al., 2014](#)). A common approach is to simply train classifiers to predict conditional probabilities at each node. Within medical imaging, hierarchical classifiers have not received much attention for CAD, but there are works on HMLC medical image retrieval ([Pourghassem and Ghasseian, 2008](#); [Demner-Fushman et al., 2015](#); [Dimitrovski et al., 2011](#)).

We present a deep HMLC approach for CXR CAD. Our work departs from prior art in three important ways. First, like other deep approaches, we train a classifier to predict conditional probabilities. However, we also demonstrate that a second fine-tuning stage, trained using unconditional probabilities, can boost performance even further. Second, we formulate a numerically stable and principled loss function for unconditional probabilities that can handle the unstable multiplication of prediction outputs. Finally, we argue that in an HMLC setting, global metrics, such as AUCs, do not provide a complete picture. Instead, we advocate also investigating performance conditioned on a high-level node being true, *e.g.*, *one or more abnormalities*, providing a measure of model performance for different patient populations, some of which may be more clinically relevant depending on the application. We evaluate our HMLC approach on the PLCO dataset ([Gohagan et al., 2000](#)), reporting a mean AUC of 0.887, the highest yet reported for this dataset. To the best of our knowledge, we are the first to outline an HMLC CAD system for medical imaging.

## 2. Methods

We introduce a two-stage method for CXR HMLC, which we first overview in Section 2.1. This is followed by Sections 2.2 and 2.3, which detail our two training stages that use conditional probability and a numerically stable unconditional probability formulation, respectively.

### 2.1. Hierarchical Multi-Label Classification

The first step in creating an HMLC system is to create the label taxonomy. Without loss of generality, we focus on the labels and data found within the CXR arm of the PLCO dataset ([Gohagan et al., 2000](#)), a large-scale lung cancer screening trial that collected structured radiological reports

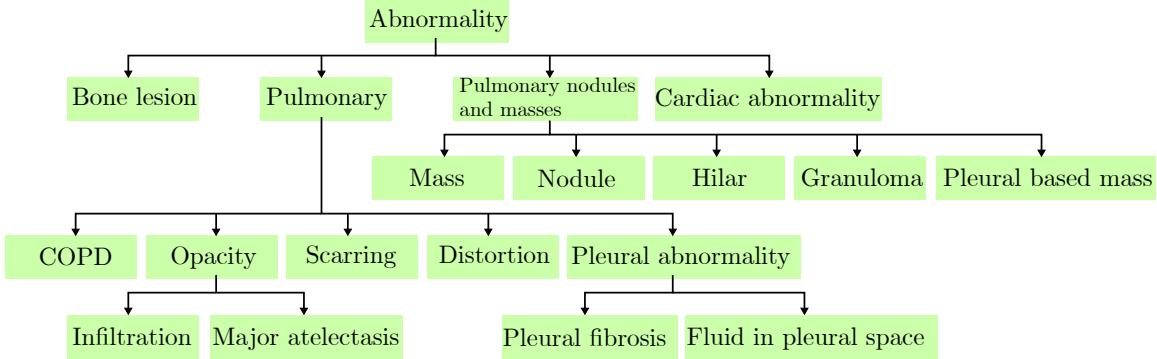


Figure 1: Constructed label hierarchy from the PLCO dataset.

of abnormalities obtained from multiple US clinical centers. From these fine-grained labels, we constructed a label taxonomy<sup>1</sup>, which is shown in Figure 1. The hierarchical structure follows the PLCO trial’s division of “suspicious for cancer” disease patterns vs. not, and is further partitioned using common groupings (Folio, 2012), totalling 19 labels. While care was taken in constructing the taxonomy and we aimed for clinical usefulness, we make no specific claim as such. We instead use the taxonomy to explore the benefits of HMLC, stressing that our approach is general enough to incorporate any appropriate taxonomy.

Because this is a multi-label setting, all or none of the labels in Figure 1 can be positive. The only restriction is that if a child is positive, its parent must be too. Siblings are not mutually exclusive. Finally, we assume that each image is associated with a set of fine-level labels and their antecedents, *i.e.*, there are no incomplete paths.

We use a DenseNet-121 (Huang et al., 2016) model as a backbone, connecting 19 fully connected layers to its last feature layer to extract 19 scalar outputs. Each output is assumed to represent the conditional probability (or its logit) given its parent is true. Thus, once the model is successfully trained, unconditional probabilities can be calculated from the output using the chain rule, *e.g.*, the unconditional probability of *scarring* can be calculated as

$$P(\text{Scarring}) = P(\text{Abnormality})P(\text{Pulmonary}|\text{Abnormality})P(\text{Scarring}|\text{Pulmonary}). \quad (1)$$

In this way, the predicted unconditional probability of a parent label is guaranteed to be greater than or equal to its children labels. We refer to the conditional probability in a label hierarchy as Hierarchical Label Conditional Probability (HLCP), and the unconditional probability calculated following the chain rule as Hierarchical Label Unconditional Probability (HLUP). The network outputs can be trained either conditionally or unconditionally, which we outline in the next two sections.

## 2.2. Training with Conditional Probability

Similar to prior work (Redmon and Farhadi, 2017; Roy et al., 2018; Yan et al., 2014), in the first stage of the proposed training scheme, each classifier is only trained on data conditioned upon its

1. Note, we merged “left hilar abnormality” and “right hilar abnormality” into “hilar abnormality”, resulting in 19 labels.

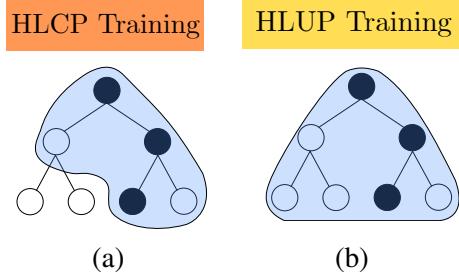


Figure 2: The HLCP and HLUP losses are depicted in (a) and (b), respectively, where black and white points are positive and negative labels, respectively. Blue areas indicate the activation area in the loss functions.

parent label being positive. Thus, training directly models the conditional probability. The shared part of the classifiers, *i.e.*, feature layers from the backbone network, is trained jointly by all the tasks. Specifically, for each image the losses are only calculated on labels whose parent label is also positive. For example, when an image with positive *Scarring* and no other positive labels is fed into training, only the losses of *Abnormality* and the children labels of *Pulmonary* and *Abnormality* are calculated and used for training.

Figure 2 (a) illustrates this training regimen, which we denote HLCP training. In this work, we use Cross Entropy (CE) loss to train the conditional probabilities, which can be written as

$$L_{HLCP} = \sum_{m \in M} CE(z_m, \hat{z}_m) * 1_{\{z_{a(m)}=1\}}, \quad (2)$$

where  $M$  denotes the set of all disease patterns, and  $m$  and  $a(m)$  denote a disease pattern and its ancestor, respectively. Here  $CE(\cdot, \cdot)$  denotes the cross entropy loss, and  $z_m \in \{0, 1\}$  denotes the ground truth label of  $m$ , with  $\hat{z}_m$  corresponding to the network’s sigmoid output.

Training with conditional probability is a very effective initialization step, as it concentrates the modeling power solely on discriminating siblings under the same parent label, rather than having to discriminate across all labels, which eases convergence and reduces confounding factors. It also alleviates the problem of low label prevalence because fewer negative samples are used for each label.

### 2.3. Fine Tuning with Unconditional Probability

In the second stage, we finetune the model using an HLUP CE loss. This stage aims at improving the accuracy of unconditional probability predictions, which is what is actually used during inference and is thus critical to classification performance. Another important advantage is that the final linear layer sees more negative samples. Predicted unconditional probabilities for label  $m$ , denoted  $\hat{p}_m$ , are calculated using the chain rule:

$$\hat{p}_m = \prod_{m' \in A(m)} \hat{z}_{m'}, \quad (3)$$

where  $A(m)$  is the union of label  $m$  and its antecedents. When training using unconditional probabilities, the loss is calculated on every classifier output for every data instance. Thus, the HLUP CE

loss for each image is simply

$$L_{HLUP} = \sum_{m \in M} CE(z_m, \hat{p}_m). \quad (4)$$

Figure 2(b) visually depicts this loss.

A naive way to calculate (4) would be a direct calculation. However, such an approach introduces instability during optimization, as the training would have to minimize the product of network outputs. In addition, the product of probability values within  $[0, 1]$  can cause arithmetic underflow. For this reason, we outline a numerically stable formulation of (4), whose derivation can be found in Appendix A:

$$L_{HLUP} = \sum_{m' \in A(m)} \ell_{m'} + \gamma, \quad (5)$$

$$\ell_{m'} = -z_m \log \left( \frac{1}{1 + \exp(-y_{m'})} \right) - (1 - z_m) \log \left( 1 - \frac{1}{1 + \exp(-y_{m'})} \right), \quad (6)$$

$$\gamma = -(1 - z_m) \left( \sum_{m' \in A(m)} y_{m'} + LSE \left( \left\{ \sum_{j \in S} -y_j \mid \forall S \in \mathcal{P}(A(m)) \setminus \{\emptyset\} \right\} \right) \right), \quad (7)$$

where  $\hat{y}_{m'}$  is the logit output for label  $m'$ . The expression in (6) is simply the CE loss given a logit input, which enjoys stable implementations within all popular deep learning software. For (7),  $\mathcal{P}(\cdot)$  denotes the powerset,  $S$  enumerates all possible subsets of  $\mathcal{P}(A(m))$ , excluding the empty set, and  $LSE(\cdot)$  is the LogSumExp function. Enumerating the powerset produces an obvious combinatorial explosion. However, for smaller-scale hierarchies, like that in Figure 1, it remains tractable. For larger hierarchies, an  $O(|A(m)|)$  solution involves simply interpreting the LogSumExp as a smooth approximation to the maximum function, but we do not need that here. Numerically stable implementations of the LogSumExp, and its gradient, are well known. Thus, since both terms in (5) can be implemented stably, our formulation avoids the numerical issues faced by a naive calculation of (4).

### 3. Experiments

We validate our approach on the PLCO dataset (Gohagan et al., 2000), which contains 198,000 manually labeled CXRs. While the recent ChestXRay14 dataset (Wang et al., 2017) is extraordinarily valuable, we expect the PLCO structured labels to have greater reliability, especially in evaluation, over the former’s text-mined labels. As noted in Section 2.1, after pre-processing the data is left with 14 leaf-node labels. We split the data into training, validation, and test sets, corresponding to 70%, 10%, and 20% of the data, respectively. Data is split at the patient level, and care was taken to balance the prevalence of each disease pattern as much as possible.

Our chosen network is DenseNet-121 (Huang et al., 2016), implemented using TensorFlow. We first train with the HLCP CE loss of (2) fine-tuning from a model pretrained from ImageNet (Deng et al., 2009). We refer to this model simply as *HLCP*. To produce our final model, we then finetune the HLCP model using the HLUP CE loss of (4). We denote this final model as *HLUP-finetune*.

While we do compare to a recent DenseNet121 BR approach (Guendel et al., 2018), we stress that direct comparisons of numbers are impossible, as Guendel et al. (2018) used different data splits and only evaluated on 12 leaf-node labels. For that reason, we also compare against three baseline

Table 1: Comparison of AUC and AP across tested models. Mean values across leaf-node and high-level disease patterns are shown, as well as leaf-node label conditioned on one or more abnormality being present.

	Leaf-node labels		High-level labels		Leaf-node labels conditioned on abnormality	
	AUC	AP	AUC	AP	AUC	AP
(Guendel et al., 2018)	0.874	N/A	N/A	N/A	N/A	N/A
BR-leaf	0.871	0.234	N/A	N/A	0.806	0.334
BR-all	0.867	0.221	0.852	0.440	0.808	0.323
HLUP	0.872	0.214	0.856	0.436	0.799	0.288
HLCP	0.879	0.229	0.857	0.440	0.822	0.329
HLUP-finetune	0.887	0.250	0.866	0.460	0.832	0.342

models, all using the same trunk network fine-tuned from ImageNet pretrained weights. The first, denoted *BR-leaf*, is trained using CE loss on the 14 leaf-node labels. This measures performance using a standard multi-label BR approach. The second, denoted *BR-all* is very similar, but trains a CE loss on all 19 labels independently, including high-level ones. In this way, *BR-all* measures performance when one wishes to naively output high-level abnormality nodes, without considering label taxonomy. Finally, we also test against a model trained using the HLUP CE loss, but not starting from the HLCP weights. As such, this baseline, denoted *HLUP*, helps reveal the impact of using a two-stage approach vs. simply training an HLUP classifier in one step. For all tested models, extensive hyper-parameter searches were performed on the NVIDIA cluster to optimize mean validation AUCs of leaf-node labels.

For all models, we evaluate the mean AUC and Average Precision (AP) on the test set. To start, we measure performance on both leaf-node as well as high-level patterns. The results are shown in the first two columns of Table 1. As the table demonstrates, the standard baseline BR-leaf model produces high AUC scores, in line with prior art (Guendel et al., 2018); however, it does not provide high-level predictions based on a taxonomy. Naively executing BR training on the entire taxonomy, *i.e.*, the BR-all model, does not improve performance. This indicates that if not properly incorporated, the label taxonomy does not benefit performance.

In contrast, the HLCP model is indeed able to match BR-leaf’s performance on the leaf-node labels, despite also being able to provide high-level predictions. HLUP-finetune goes further by exceeding BR-leaf’s performance, demonstrating that our two-stage training process can produce tangible improvements. This is underscored when comparing HLUP-finetune with HLUP, which highlights that without the two-stage training, HLUP training cannot reach the same performance. If we limit ourselves to models incorporating the entire taxonomy, our final HLUP-finetune model outperforms BR-all by 2% and 2.9% in leaf-node mean AUC and AP values, respectively. Figure 3 provides more details on these improvements, demonstrating that AUC values are higher for HLUP-finetune compared to the baseline method for all leaf-node and high-level disease patterns. Although not graphed here for clarity reasons, HLUP-finetune also outperformed the HLCP method for all

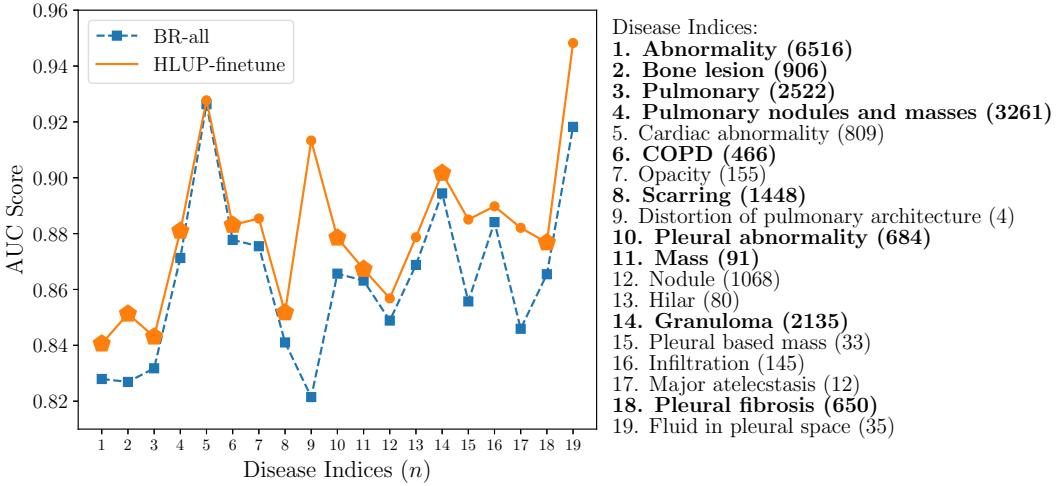


Figure 3: Comparison of AUC scores for all leaf-node and high-level disease patterns for the BR-all and HLUP-finetune models. The dashed line separates the leaf-node from the high-level disease patterns. Bolded labels and larger graph markers, denote disease patterns exhibiting statistically significant improvement ( $p < 0.05$ ) using the StAR software implementation (Vergara et al., 2008) of the non-parametric test of DeLong et al. (1988).

disease patterns. Of note, is that significance values also respect the disease hierarchy, and if a child disease pattern demonstrates statistically significant improvement, so does its parent.

Because more than one label can be positive, multi-label classification performance has exponentially more facets for evaluation than single-label or even multi-class settings. Here, we explore one such facet, namely model performance conditioned on high-level nodes being positive. We restrict our focus to CXRs exhibiting one or more disease patterns, *i.e.*, *abnormality* being positive. As such, this sheds light on model performance when it may be critical to discriminate what combination of disease patterns are present, which is crucial for proper CXR interpretation (Folio, 2012). The last column of Table 1 depicts these results. As can be seen, in such settings, HLUP-finetune still exhibits increased performance over the baseline models and also the next-best hierarchical model. Importantly, if we compare the conditional AUCs between BR-all and HLUP-finetune, we see a 2.4% increase. As a result, in the critical setting of a CXR exhibiting at least one disease pattern, our HLUP-finetune still manages to provide key performance improvements.

Finally, we compare our numerically stable implementation of HLUP CE loss in (5) to: (a) the naive approach of directly optimizing (3); and (b) to a recent rescaling approximation, originally introduced for the multiplication of independent, rather than conditional probabilities, seen in multi-instance learning (Li et al., 2018). This latter approach rescales each individual probability multiplicand in (3) to guarantee that the product is greater than or equal to 1e-7. Similar to the naive approach, the product is then optimized directly using CE loss. Based on a maximum depth of four for our taxonomy, we implement this approach by rescaling each multiplicand in (3) to [0.02, 1]. As Table 2 demonstrates, regardless if we train from ImageNet or finetune from the HLCP model, our numerically stable formulation far outperforms this rescaling approximation. However, while our

Table 2: Comparison of AUCs produced using different HLUP CE loss implementations.

HLUP (naive)	HLUP (rescale)	HLUP (ours)	HLUP- finetune (naive)	HLUP- finetune (rescale)	HLUP- finetune (ours)
0.864	0.853	0.872	0.886	0.867	0.887

HLUP loss outperforms the naive implementation when training from ImageNet weights, it does not exhibit improvements when fine-tuning from the HLCM model. We hypothesize that the predictions for the HLCM are already at a good enough quality that the numerical instabilities of the naive HLUP CE loss are not severe enough to impair performance. Nonetheless, given the improvements when training from ImageNet weights, these results indicate that our HLCM CE loss does indeed provide tangible improvements in convergence stability. We expect these improvements to be greater given taxonomies of greater depth, and our formulation should also prove valuable to multi-instance setups which must optimize CE loss over the product of large numbers of probabilities, *e.g.*, the 256 multiplicands seen in [Li et al. \(2018\)](#).

## 4. Conclusion

We have presented a two-stage approach for deep HMLC of CXRs that combines conditional training with an unconditional probability fine-tuning step. To effect the latter, we introduce a new and numerically stable formulation for HLUP CE loss, which we expect would also prove valuable in other training scenarios involving the multiplication of probability predictions, *e.g.*, multi-instance learning. Through comprehensive evaluations, we report the highest yet mean AUC on the PLCO dataset, outperforming hierarchical and non-hierarchical alternatives. We also show performance improvements conditioned on one or more abnormalities being present, *i.e.*, predicting the specific combination of disease patterns, which is crucial for CXR interpretation. Experiments also demonstrate that HLUP fine-tuning is crucial in achieving these results. Future work should focus on characterizing improvements against the recently released CheXpert dataset ([Irvin et al., 2019](#)) and also on computer vision benchmarks. Additionally, another potential strength of the HMLC approach is handling incomplete labels, which also deserves further investigation. Finally, another interesting focus would be exploring whether using hierarchical features, rather than the shared ones of our approach, would improve results further.

## 5. Acknowledgements

We thank the National Cancer Institute (NCI) for access to NCI’s data collected by the PLCO Cancer Screening Trial. The statements contained herein are solely ours and do not represent or imply concurrence or endorsement by NCI. We also thank Chaochao Yan for help on pre-processing the PLCO images and labels.

## References

- W. Bi and J. T. Kwok. Bayes-optimal hierarchical multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):2907–2918, Nov 2015. ISSN 1041-4347. doi: 10.1109/TKDE.2015.2441707.
- Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837–845, 1988.
- Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Mach. Learn.*, 88(1-2):5–45, July 2012. ISSN 0885-6125. doi: 10.1007/s10994-012-5285-8. URL <https://doi.org/10.1007/s10994-012-5285-8>.
- Dina Demner-Fushman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, and George R. Thoma. Annotation of chest radiology reports for indexing and retrieval. In Henning Müller, Oscar Alfonso Jimenez del Toro, Allan Hanbury, Georg Langs, and Antonio Foncubierta Rodriguez, editors, *Multimodal Retrieval in the Medical Domain*, pages 99–111, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24471-6.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sašo Deroski. Hierarchical annotation of medical images. *Pattern Recogn.*, 44(10-11):2436–2449, October 2011. ISSN 0031-3203. doi: 10.1016/j.patcog.2011.03.026. URL <http://dx.doi.org/10.1016/j.patcog.2011.03.026>.
- Les Folio. *Chest imaging: An algorithmic approach to learning*. Springer, 01 2012.
- John K. Gohagan, Philip C. Prorok, Richard B. Hayes, and Barnett-S. Kramer. The prostate, lung, colorectal and ovarian (plco) cancer screening trial of the national cancer institute: History, organization, and status. *Controlled Clinical Trials*, 21(6, Supplement 1):251S – 272S, 2000. ISSN 0197-2456. doi: [https://doi.org/10.1016/S0197-2456\(00\)00097-0](https://doi.org/10.1016/S0197-2456(00)00097-0). URL <http://www.sciencedirect.com/science/article/pii/S0197245600000970>.
- Sebastian Guendel, Sasa Grbic, Bogdan Georgescu, Kevin Zhou, Ludwig Ritschl, Andreas Meier, and Dorin Comaniciu. Learning to recognize abnormalities in chest x-rays with location-aware dense networks, 2018. arXiv:1803.04565.
- Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. URL <http://arxiv.org/abs/1608.06993>.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Illcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019. URL <http://arxiv.org/abs/1901.07031>.

Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Jenifer Siegelman, Les Folio, Sameer Antani, and George Thoma. Automatic screening for tuberculosis in chest radiographs: a survey. *Quantitative Imaging in Medicine and Surgery*, 3(2):89–99, April 2013. ISSN 2223-4292. doi: 10.3978/j.issn.2223-4292.2013.04.03.

Curtis P. Langlotz. Radlex: A new method for indexing online educational materials. *RadioGraphics*, 26(6):1595–1597, 2006. doi: 10.1148/rg.266065168. URL <https://doi.org/10.1148/rg.266065168>. PMID: 17102038.

Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8290–8299, 2018.

Hossein Pourghassem and Hassan Ghassemian. Content-based medical image classification using a new hierarchical merging scheme. *Computerized Medical Imaging and Graphics*, 32(8):651 – 661, 2008.

J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 6517–6525, July 2017. doi: 10.1109/CVPR.2017.690. URL [doi.ieee.org/10.1109/CVPR.2017.690](http://doi.ieee.org/10.1109/CVPR.2017.690).

Deboleena Roy, Priyadarshini Panda, and Kaushik Roy. Tree-cnn: A deep convolutional neural network for lifelong learning. *CoRR*, abs/1802.05800, 2018. URL <http://arxiv.org/abs/1802.05800>.

Ismael A Vergara, Tomás Norambuena, Evandro Ferrada, Alex W Slater, and Francisco Melo. StAR: a simple tool for the statistical comparison of ROC curves. *BMC bioinformatics*, 9:265–265, June 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-265. URL <https://www.ncbi.nlm.nih.gov/pubmed/18534022>.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammad Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CoRR*, abs/1705.02315, 2017. URL <http://arxiv.org/abs/1705.02315>.

Zhicheng Yan, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Robinson Piramuthu. HD-CNN: hierarchical deep convolutional neural network for image classification. *CoRR*, abs/1410.0736, 2014. URL <http://arxiv.org/abs/1410.0736>.

Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels, 2017. arXiv:1710.10501.

## Appendix A. Numerically Stable Formulation of HLUP CE Loss

Denoting the network's output logits as  $\hat{y}_{(.)}$ , the predicted unconditional probability of label  $m$  can be written as:

$$\hat{p}_m = \prod_{m'} \frac{1}{1 + \exp(-y_{m'})}, \quad (8)$$

where we use  $m'$  to denote  $m' \in A(m)$  for notational simplicity.

The HLUP CE loss is calculated as:

$$L_{HLUP} = -z_m \log(\hat{p}_m) - (1 - z_m) \log(1 - \hat{p}_m), \quad (9)$$

$$= -z_m \log \left( \prod_{m'} \frac{1}{1 + \exp(-y_{m'})} \right) - (1 - z_m) \log \left( 1 - \left( \prod_{m'} \frac{1}{1 + \exp(-y_{m'})} \right) \right), \quad (10)$$

where  $z_m$  is the ground truth label of  $m$ .

We would like to break up the second term in (10) to produce the following formulation:

$$L_{HLUP} = -z_m \log \left( \prod_{m'} \frac{1}{1 + \exp(-y_{m'})} \right) - (1 - z_m) \log \left( \prod_{m'} \left( 1 - \frac{1}{1 + \exp(-y_{m'})} \right) \right) + \gamma \quad (11)$$

$$= \sum_{m'} \left( -z_m \log \left( \frac{1}{1 + \exp(-y_{m'})} \right) - (1 - z_m) \log \left( 1 - \frac{1}{1 + \exp(-y_{m'})} \right) \right) + \gamma, \quad (12)$$

which can be simplified to a sum of individual CE losses plus  $\gamma$ :

$$L_{HLUP} = \sum_{m'} \ell_{m'} + \gamma, \quad (13)$$

where  $\ell_m$  are individual cross entropy terms, using  $z_m$  and  $y_{m'}$  as the ground truth and logit input, respectively, and  $\gamma$  is the scalar quantity we want to formulate. Note that (13) allows us to take advantage of numerically stable CE implementations, *e.g.*, those within Tensorflow, to calculate  $\sum_{m'} \ell_{m'}$ .

To satisfy (12), we will need  $\gamma$  to satisfy:

$$\begin{aligned} -(1 - z_m) \log \left( \prod_{m'} \left( 1 - \frac{1}{1 + \exp(-y_{m'})} \right) \right) + \gamma &= -(1 - z_m) \log \left( 1 - \left( \prod_{m'} \frac{1}{1 + \exp(-y_{m'})} \right) \right), \\ \log \left( \prod_{m'} \left( 1 - \frac{1}{1 + \exp(-y_{m'})} \right) \right) - \frac{\gamma}{1 - z_m} &= \log \left( 1 - \left( \prod_{m'} \frac{1}{1 + \exp(-y_{m'})} \right) \right), \\ \left( \prod_{m'} \left( 1 - \frac{1}{1 + \exp(-y_{m'})} \right) \right) \exp \left( -\frac{\gamma}{1 - z_m} \right) &= 1 - \left( \prod_{m'} \frac{1}{1 + \exp(-y_{m'})} \right). \end{aligned} \quad (14)$$

Denoting

$$\alpha = \exp \left( -\frac{\gamma}{1 - z_m} \right), \quad (15)$$

we have:

$$\alpha \left( \prod_{m'} \left( 1 - \frac{1}{1 + \exp(-y_{m'})} \right) \right) = 1 - \prod_{m'} \frac{1}{1 + \exp(-y_{m'})}, \quad (16)$$

$$\alpha \left( \frac{\prod_{m'} \exp(-y_{m'})}{\prod_{m'} (1 + \exp(-y_{m'}))} \right) = \frac{\prod_{m'} (1 + \exp(-y_{m'})) - 1}{\prod_{m'} (1 + \exp(-y_{m'}))}, \quad (17)$$

$$\alpha = \frac{\prod_{m'} (1 + \exp(-y_{m'})) - 1}{\exp(\sum_{m'} -y_{m'})}. \quad (18)$$

Substituting the left side of (18) into (15) gives us:

$$\begin{aligned} \gamma &= -(1 - z_m) \log(\alpha) \\ &= -(1 - z_m) \left( \sum_{m'} y_{m'} + \log \left( \prod_{m'} (1 + \exp(-y_{m'})) - 1 \right) \right). \end{aligned} \quad (19)$$

If the log-product term of (19) is expanded, with 1 subtracted, it will result in

$$\gamma = -(1 - z_m) \left( \sum_{m'} y_{m'} + \log \left( \sum_{S \in \mathcal{P}(A(m)) \setminus \{\emptyset\}} \exp \left( \sum_{j \in S} -y_j \right) \right) \right), \quad (20)$$

where  $S$  enumerates all possible subsets of the powerset of  $A(m)$ , excluding the empty set. For example if there were two logits,  $y_1$  and  $y_2$ , the summation inside the log would be:

$$\exp(-y_1) + \exp(-y_2) + \exp(-y_1 - y_2). \quad (21)$$

The expression in (20) can be written as

$$\gamma = -(1 - z_m) \left( \sum_{m'} y_{m'} + LSE \left( \left\{ \sum_{j \in S} -y_j \quad \forall S \in \mathcal{P}(A(m)) \setminus \{\emptyset\} \right\} \right) \right), \quad (22)$$

where  $LSE$  is the LogSumExp function. Many numerical packages, including TensorFlow, provide numerically stable implementations of  $LSE$ , and its derivative. By substituting (22) into (12), a numerically stable version of the HLUP CE loss can be calculated.

Should the cardinality of the powerset be too high, the LogSumExp expression can be approximated as a maximum function, which can be calculated using an  $O(|A(m)|)$  scan of  $y_{m'}$  values:

$$\gamma \approx -(1 - z_m) \left( \sum_{m'} y_{m'} + \max \left( \left\{ \sum_{j \in S} -y_j \quad \forall S \in \mathcal{P}(A(m)) \setminus \{\emptyset\} \right\} \right) \right), \quad (23)$$

$$= \begin{cases} -(1 - z_m) \left( \sum_{m'} y_{m'} + \sum_{j: y_j < 0} -y_j \right), & \text{if } \exists y_{m'} < 0 \\ -(1 - z_m) (\sum_{m'} y_{m'} + \max(\{-y_{m'}\})), & \text{otherwise} \end{cases}. \quad (24)$$

# Digitally Stained Confocal Microscopy through Deep Learning

**Marc Combalia<sup>1</sup>**

MCOMBALIA@CLINIC.CAT

**Javiera Pérez-Anker<sup>1</sup>**

PEREZ12@CLINIC.CAT

**Adriana García-Herrera<sup>2</sup>**

APGARCIA@CLINIC.CAT

**Llúcia Alos<sup>2</sup>**

LALOS@CLINIC.CAT

**Verónica Vilaplana<sup>3</sup>**

VERONICA.VILAPLANA@UPC.EDU

**Ferran Marqués<sup>3</sup>**

FERRAN.MARQUES@UPC.EDU

**Susana Puig<sup>1</sup>**

SPUIG@CLINIC.CAT

**Josep Malvehy<sup>1</sup>**

JMALVEHY@CLINIC.CAT

<sup>1</sup> Dermatology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Barcelona, Spain.

<sup>2</sup> Pathology Department, Melanoma Unit, Hospital Clínic de Barcelona, IDIBAPS, Universitat de Barcelona, Barcelona, Spain.

<sup>3</sup> Signal Theory and Communications Department, Universitat Politècnica de Catalunya. BarcelonaTech, Spain.

## Abstract

Specialists have used confocal microscopy in the ex-vivo modality to identify Basal Cell Carcinoma tumors with an overall sensitivity of 96.6% and specificity of 89.2% (Chung et al., 2004). However, this technology hasn't established yet in the standard clinical practice because most pathologists lack the knowledge to interpret its output. In this paper we propose a combination of deep learning and computer vision techniques to digitally stain confocal microscopy images into H&E-like slides, enabling pathologists to interpret these images without specific training. We use a fully convolutional neural network with a multiplicative residual connection to denoise the confocal microscopy images, and then stain them using a Cycle Consistency Generative Adversarial Network.

**Keywords:** Deep learning, Neural Networks, Digital Staining, Confocal Microscopy, Speckle Noise, CycleGAN

## 1. Introduction

Histopathology with hematoxylin and eosin (H&E) staining is widely used as a diagnostic tool for a large variety of tissue lesions. However, it requires skilled technicians to process and stain the tissue samples, and it is very costly and time-consuming, requiring from hours to days before a pathologist can analyze the samples. These long delays often impede rapid evaluation of lesions during a surgical operation.

Confocal microscopy (CM) is a novel technique for tissue examination where a laser is focused on a microscopic target and the scattering of the light through its various structures is captured to form a two-dimensional grayscale image (Calzavara-Pinton et al., 2008). These microscopes can operate in two different modes (reflectance (RCM) and fluorescence (FCM)) which highlight different microscopic structures in the tissue. The combination of the two modes can improve the

diagnostic accuracy of the pathologist in the ex-vivo evaluation of tumour margins ([Gareau, 2009](#)). In the last years, this new technology has enabled the rapid evaluation of tissue samples directly in the surgery room significantly reducing the time of complex surgical operations in skin cancer ([Cinotti et al., 2018](#)).

CMs can obtain images with an optical resolution comparable to pathology, but their output largely differs from the standard H&E slides that pathologists use to evaluate in their clinical practise. Some researchers have focused on creating digitally stained (H&E)-like images from the output of the CMs to facilitate their interpretation by untrained pathologists and surgeons. ([Gareau, 2009](#)) has proposed a digital staining technique which linearly combines the FCM and RCM images to form an RGB output slide which resembles H&E stained pathology giving a blue color to FCM and pink to purple color to RCM. This is, in fact, the algorithm used in the last generation of the Vivascope 2500 clinical ex vivo CM device ([Vivascope, 2018](#)). This simple staining technique is good at enhancing cellular details allowing the mitosis visualization, but its colors and structures greatly vary from the ones found in the original H&E slides.

In this work, we propose a deep learning technique to combine the two modes of the CM into a (H&E)-like image. First, a fully convolutional neural network is used to remove the speckle noise present in the RCM images ([Wang et al., 2017](#)), and then a Cycle Consistency Generative Adversarial Network (CycleGAN) ([Zhu et al., 2017](#)) is used to combine the FCM and RCM modes into a digitally stained (H&E) slide.

## 2. Materials and Methods

In this section, we describe the architecture of the Despeckling Neural Network used in the RCM image of the CM and the Generative Adversarial Network used to create the (H&E)-like digitally stained image. Figure 1 shows the complete pipeline for CM image staining.

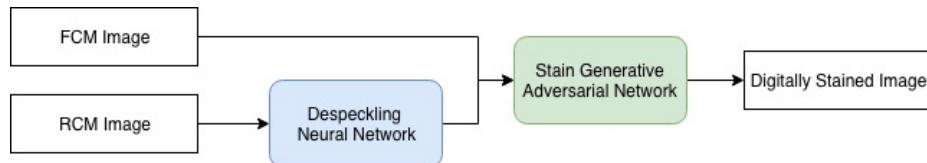


Figure 1: Diagram of the proposed architecture to transform the output of the CM to digitally stained (H&E)-like slides.

### 2.1. Reflectance Image Despeckling

RCM images are corrupted with a kind of multiplicative noise known as speckle ([Sarode and Deshmukh, 2011](#)). Speckle noise is due to a combination of constructive and destructive fluctuations at the input of the CM sensor which interfere with the nominal tissue structure reflectance. The presence of speckle noise limits the application of further post-processing and computer vision techniques and makes diagnosing less reliable for physicians ([Gigilashvili, 2017](#)). Hence, before digitally staining the CM images, their noise must be reduced. Figure 2 shows some RCM images extracted from the CM dataset presented in section [2.3.1](#).

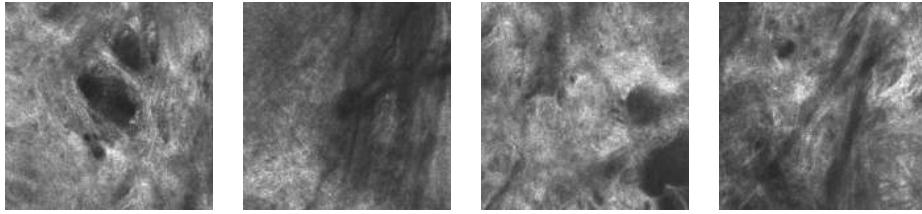


Figure 2: Reflectance images with speckle noise at the output of the CM.

In this work, similarly to (Wang et al., 2017), we use a fully convolutional neural network together with a residual connection to reduce the intensity of the noise present in the RCM images. The observed image at the output of the CM is related to real tissue reflectance by the following equation:

$$Y = X * (1 + F)$$

Where  $Y \in \mathcal{R}^{W \times H}$  is the observed RCM image,  $X \in \mathcal{R}^{W \times H}$  is the noise-free reflectance of the tissue, and  $F \in \mathcal{R}^{W \times H}$  is the speckle noise random variable. We include the aforementioned formulation inside the architecture of the neural network so that it is trained to estimate the inverse of the speckle noise  $1/\hat{F}$  at the last convolutional layer.

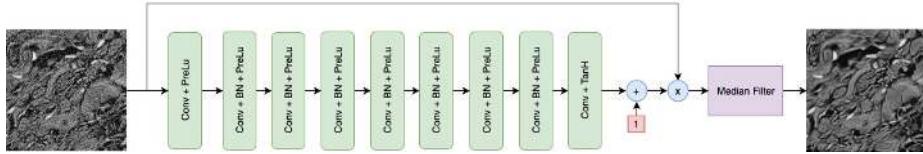


Figure 3: Architecture of the despeckling neural network used in our experiments.

The fully convolutional neural network is composed of 7 convolutional layers (PreLu activations) with 64 filters each and 1 final convolutional layer (Tanh activation) with a single filter. Batch normalization is added to the intermediate layers of the neural network. We use a multiplicative residual connection between the last convolutional layer and the input image to incorporate the speckle noise formulation in the the neural network. The architecture is trained to minimize the squared error between the clean images and its output. After training, some noise may still be present in some isolated pixels (figure 6). Authors in (Wang et al., 2017) add total variation loss to the training process to remove these spurious pixel activations. Instead, we filter out the remaining noise using a 3x3 median filter. We train the neural network on a dataset of skin histology images which have been artificially contaminated, which have a similar appearance to noisy RCM images.

## 2.2. Confocal Microscopy Staining

Due to the impossibility to obtain paired data between the CM domain (A) and the stained H&E histology domain (B) (tissue blocks scanned with the CM need to undergo slicing before staining

with H&E), we use Cycle Consistency Generative Adversarial Networks (CycleGAN) (Zhu et al., 2017) to transfer the H&E stain appearance to the CM images. The CycleGAN architecture consists of two generator and discriminator pairs. The first pair tries to map images from domain A to domain B, while the second pair undergoes the contrary operation. The generators’ task is to create images that the discriminators can’t distinguish from real samples. We use a ResNet (He et al., 2016) architecture in the generators and a PatchNet (Isola et al., 2016) in the discriminators. Figure 4 shows all the components and loss functions involved in the translation from A to B.

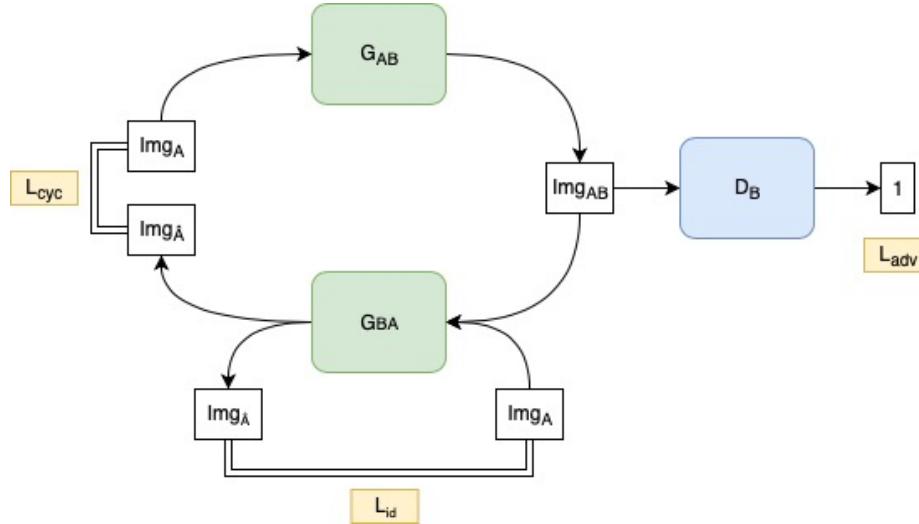


Figure 4: Components and loss functions ( $L_{cyc}$ ,  $L_{adv}$ ,  $L_{id}$ ) involved in the domain translation from A to B. The same process is carried out on the contrary direction when translating from B to A.

It is known that CycleGANs are sensitive to their initialization, so to pose an easier task, we use the digital staining method proposed in (Gareau, 2009) as source images for domain A. Figure 5 shows this transformation.

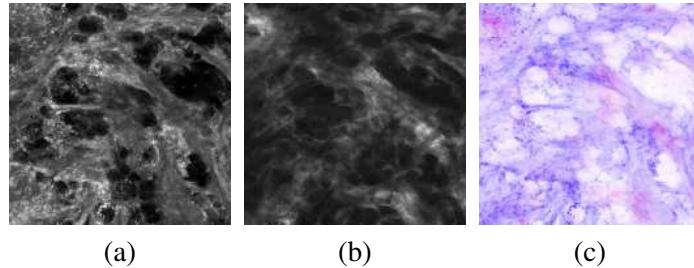


Figure 5: (a) FCM image, (b) RCM image, (c) Digital staining as proposed by (Gareau, 2009)

## 2.3. Data

### 2.3.1. CONFOCAL MICROSCOPY

Our CM dataset consists of 11 microscopy skin slides obtained with the Vivascope 2500 4th Generation CM ([Vivascope, 2018](#)), which captures the tissues at a resolution of  $0.75 \mu m/px$ . Its output consists of two large grayscale images (more than 10000 x 10000 pixels), one for each capture mode (RCM and FCM). Both modes are normalized to cover a range from 0 to 1, and OTSU thresholding is used to determine the tissue-containing regions. Non-overlapping patches of size 1024 x 1024 pixels are extracted summing a total of 949 1024x1024 images for each mode. The patches are divided into train (80%) and validation (20 %) taking into account their origin slide.

### 2.3.2. H&E HISTOLOGY

Our H&E Histology dataset consists of 29 skin tissue samples obtained with a Ventana Whole Slide Image scanner captured with a resolution of  $0.47 \mu m/px$ . OTSU thresholding is used to determine the tissue containing regions in each whole slide image and multiple patches of size 1630 x 1630 pixels are extracted and then resized to 1024x1024 pixels. The final resolution of each patch is the same as the CM resolution ( $0.75 \mu m/px$ ). The processed dataset consists of a total of 8789 images (80 % for the training split and 20 % for the validation split, partitioned taking into account their origin slide).

## 3. Experiments and Results

In this section, we describe the results obtained for the Despeckling Neural Network used in the RCM of the CM and the Generative Adversarial Network used to create the (H&E)-like digitally stained image.

### 3.1. Reflectance Despeckling

Since it is not possible to obtain noise-free reflectance images at the output of the CM to train the neural network, we chose to train the despeckling neural network on histology images since the structures present in both domains are the same ([Ragazzi et al., 2014](#)). The main differences between reflectance mode confocal images and histology images are two: reflectance images are one channel only, and tissue structures appear lighter than the background. To create the dataset to train the neural network, we transformed the histology images to the YUV color space and used to the inverse of the Y channel to train the despeckling neural network (with artificial speckle noise;  $noisy_{xy} = clean_{xy} * noise$  where  $noise = Gauss(mean = 1, std = 0.2)$ ). We trained the Despeckling Neural network on 7031 histology images. We augmented the training dataset through the use of random flips and random crops of size of 256 x 256 pixels and we updated the weights of the neural network using Adam optimization with a learning rate of  $5e - 4$ . Finally, we evaluated the results on 1758 histology images contaminated with artificial speckle noise and 949 RCM images extracted from the CM. In Table 1 we present the quantitative results obtained on the artificial dataset, and figures 6 and 7 show some images before and after going through the despeckling neural network for the artificial histology dataset and RCM dataset respectively.

Table 1: PSNR and SSIM before and after applying the proposed Despeckling Neural Network.

Error Measure	Noisy	Despeckled
PSNR (dB)	16.19	23.97
SSIM	0.438	0.727

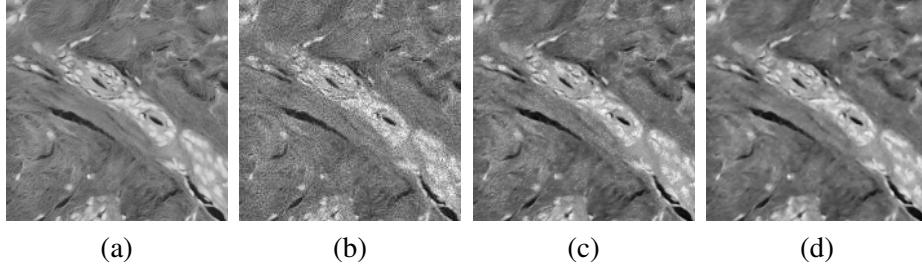


Figure 6: (a) Original clean image from the artificial despeckling dataset, (b) Original image with artificial speckle noise, (c) Despeckled image at the output of the neural network, before the 3x3 median filter, (d) Despeckled image after the 3x3 median filter.

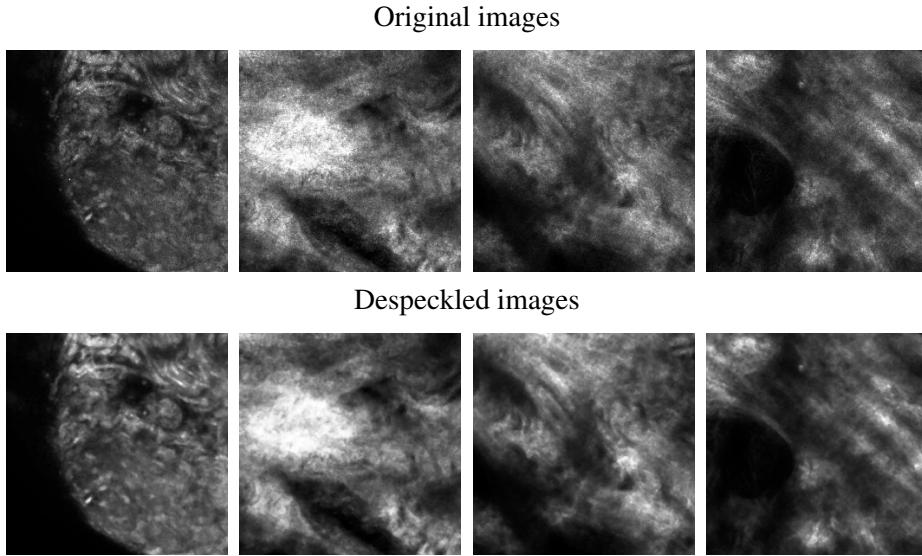


Figure 7: RCM images before and after the proposed despeckling neural network.

### 3.2. Confocal Image Staining

We trained the Staining CycleGAN on 759 CM images and 282 histology images extracted from a single slide (the one with the best proportion of hematoxylin and eosin stains). The CycleGAN was trained with Adam optimization and a learning rate of  $2e - 4$ . We trained the neural network first on patches of size 256 x 256 pixels. After 50 epochs, we augmented the patch size to 512 x 512 so that the architecture could learn features seen at a higher scale and then trained it for another 50 epochs with learning rate decay. Figure 8 shows some results of a CycleGAN trained on images which

have been previously denoised with the method described in section 2.1, as well as some images from domain B. Figure 9 shows some results of a CycleGAN trained with noisy RCM images. The trained architecture can digitally stain a 15000 x 10000 pixels confocal microscopy slide in less than 3 minutes using a NVIDIA Tesla K80.

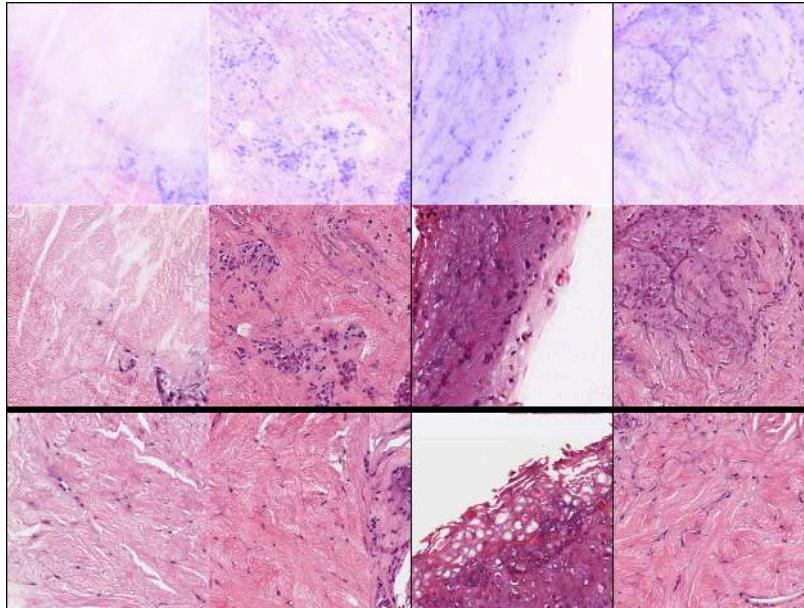


Figure 8: Results of the proposed architecture (Despeckling neural network and CycleGAN). The top row represents the input images of the CycleGAN, which have been digitally stained with the method proposed by (Gareau, 2009). Middle row is the output of the staining CycleGAN. The images in the bottom row are real H&E stained histology images extracted from the training dataset. All images are 512x512 pixels.

#### 4. Discussion

We argue that the combination of the proposed despeckling neural network with the CycleGAN architecture for stain transfer is capable of producing realistic (H&E)-like images. Output images from the proposed algorithm were evaluated by two expert pathologists in our department (LL.A/A.G) and they confirmed that the images were similar to those in routine.

The despeckling neural network was able to successfully remove the noise from the RCM images at the output of the CM. From the results on figure 9 we conclude that the Despeckling Neural Network is crucial to obtain realistic images at the output of the CycleGAN. The architecture trained with noisy RCM images had a harder time learning to map the confocal output to the (H&E)-like appearance and produced non-desirable artifacts, as well as eliminated some nuclei present in the CM images. However, we argue that the Despeckling Neural Network could benefit from including an adversarial loss in the optimization process to produce sharper results.

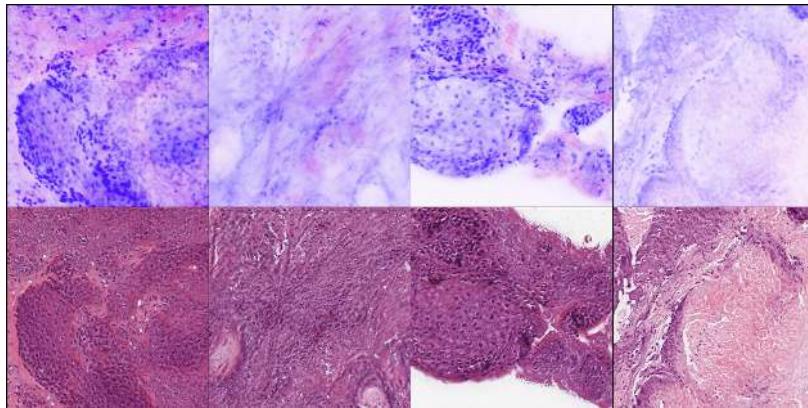


Figure 9: Results from the CycleGAN trained with RCM noisy images. Top row represents the input images of the CycleGAN, which have been digitally stained with the method proposed by (Gareau, 2009). Bottom row is the output of the staining CycleGAN. All images are 512x512 pixels.

## 5. Conclusions and Future Work

We have proposed an architecture which successfully addresses the problems involved in CM image staining. On the one hand, we reduce the noise present in the RCM images through the use of a denoising convolutional neural network with a multiplicative residual connection. Then, the denoised RCM image and FCM image are combined in a generative adversarial network to produce a realistic (H&E)-like output image. The methods described in this paper will undergo clinical validation and their diagnostic accuracy will be tested in the near future.

## References

- Piergiacomo Calzavara-Pinton, Caterina Longo, Marina Venturini, Raffaella Sala, and Giovanni Pellacani. Reflectance confocal microscopy for in vivo skin imaging. *Photochemistry and photobiology*, 84(6):1421–1430, 2008.
- Vinh Q Chung, Peter J Dwyer, Kishwer S Nehal, Milind Rajadhyaksha, Gregg M Menaker, Carlos Charles, and S Brian Jiang. Use of ex vivo confocal scanning laser microscopy during mohs surgery for nonmelanoma skin cancers. *Dermatologic surgery*, 30(12p1):1470–1478, 2004.
- Elisa Cinotti, Jean Luc Perrot, Bruno Labeille, Frédéric Cambazard, and Pietro Rubegni. Ex vivo confocal microscopy: an emerging technique in dermatology. *Dermatology practical & conceptual*, 8(2):109, 2018.
- Daniel S Gareau. Feasibility of digitally stained multimodal confocal mosaics to simulate histopathology. *Journal of biomedical optics*, 14(3):034050, 2009.
- Davit Gigilashvili. Measuring and mitigating speckle noise in dual-axis confocal microscopy images. Master’s thesis, NTNU, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. URL <http://arxiv.org/abs/1611.07004>.

Moira Ragazzi, Simonetta Piana, Caterina Longo, Fabio Castagnetti, Monica Foroni, Guglielmo Ferrari, Giorgio Gardini, and Giovanni Pellacani. Fluorescence confocal microscopy for pathologists. *Modern Pathology*, 27(3):460, 2014.

M.V. Sarode and P.R. Deshmukh. Reduction of speckle noise and image enhancement of images using filtering technique. *International Journal of Advancements in Technology*, 2:30–38, 01 2011.

Vivascope. Vivascope. <http://www.vivascope.de/home.html>, 2018. Accessed: 2018-12-13.

Puyang Wang, He Zhang, and Vishal M Patel. Sar image despeckling using a convolutional neural network. *IEEE Signal Processing Letters*, 24(12):1763–1767, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. URL <http://arxiv.org/abs/1703.10593>.

# Deep Reinforcement Learning for Subpixel Neural Tracking

**Tianhong Dai<sup>1</sup>**

**Magda Dubois<sup>1</sup>**

**Kai Arulkumaran<sup>1</sup>**

**Jonathan Campbell<sup>1</sup>**

**Cher Bass<sup>1</sup>**

**Benjamin Billot<sup>1</sup>**

**Fatmatulzehra Uslu<sup>1</sup>**

**Vincenzo de Paola<sup>2</sup>**

**Claudia Clopath<sup>1</sup>**

**Anil Anthony Bharath<sup>1</sup>**

TIANHONG.DAI15@IMPERIAL.AC.UK

MAGDA.DUBOIS.18@UCL.AC.UK

KAILASH.ARULKUMARAN13@IMPERIAL.AC.UK

JONATHAN.CAMPBELL13@IMPERIAL.AC.UK

C.BASS14@IMPERIAL.AC.UK

BENJAMIN.BILLOT.18@UCL.AC.UK

F.USLU13@IMPERIAL.AC.UK

VINCENZO.DEPAOLA@CSC.MRC.AC.UK

C.CLOPATH@IMPERIAL.AC.UK

A.BHARATH@IMPERIAL.AC.UK

<sup>1</sup> Department of Bioengineering, Imperial College London, London, UK

<sup>2</sup> MRC Clinical Science Centre, Faculty of Medicine, Imperial College London, London, UK

## Abstract

Automatically tracing elongated structures, such as axons and blood vessels, is a challenging problem in the field of biomedical imaging, but one with many downstream applications. Real, labelled data is sparse, and existing algorithms either lack robustness to different datasets, or otherwise require significant manual tuning. Here, we instead learn a tracking algorithm in a synthetic environment, and apply it to tracing axons. To do so, we formulate tracking as a reinforcement learning problem, and apply deep reinforcement learning techniques with a continuous action space to learn how to track at the subpixel level. We train our model on simple synthetic data and test it on mouse cortical two-photon microscopy images. Despite the domain gap, our model approaches the performance of a heavily engineered tracker from a standard analysis suite for neuronal microscopy. We show that fine-tuning on real data improves performance, allowing better transfer when real labelled data is available. Finally, we demonstrate that our model’s uncertainty measure—a feature lacking in hand-engineered trackers—corresponds with how well it tracks the structure.

**Keywords:** tracking, tracing, neuron, axon, reinforcement learning, transfer learning

## 1. Introduction

Although image segmentation has received significant attention as a tool for analysing biomedical image data (Greenspan et al., 2016), it does not immediately provide geometric information. Indeed, semantic pixel-level segmentation is often an input to further analytical processes: measuring sizes, areas, or being used as inputs to global shape representations. In contrast, tracking—often implemented through Kalman filtering, particle filtering, or semi-heuristic connectivity algorithms, can provide additional structural information. In particular, tracking differs from segmentation in several crucial ways:

- Tracking establishes an *order* to locations;
- Tracking can be used to capture semantically useful properties of data (e.g. velocity of movement; length of a structure) without an explicit label being applied to each observation point;

- Algorithms for tracking (Bar-Shalom and Li, 1995; Van Trees and Bell, 2007) often include some form of model-based parameter estimation for properties of interest (e.g., Kalman filters for velocity estimation).

Tracking may also involve solving some form of correspondence, or target assignment problem: this is particularly true when multiple structures are being tracked, i.e., multiple, distinct paths exist within an image, and also when objects have gaps due to geometric factors associated with slice selection or confocal imaging. An analogy can be drawn to tracking pedestrians from video data: multiple pedestrians may need to be tracked, and occlusions from other people or objects are common. While a segmentation of an image frame would typically exclude occluded objects, a tracking algorithm would need to infer the objects' locations through time and space.

Tracking is of interest in biomedical imaging where it can be applied to analyse thin, elongated structures that might vary in apparent contrast and curvature, or have crossing and branching patterns. The structured output that can be produced by tracking is useful to quantify different aspects of the underlying geometry (e.g., number of structures/branches, crossing point locations, etc.), which is not possible with segmentation. As such, tracking algorithms have been applied to biomedical datasets with thin structures, for example: neurons (Meijering, 2010; Peng et al., 2010, 2015; Acciai et al., 2016; Poulin et al., 2017), blood vessels (Fraz et al., 2012; Kumar et al., 2015), and muscle fibres (Farris and Lichtwardt, 2016).

To alleviate the need for hand-engineering trackers for different biomedical image datasets, we first formulate the task of tracing paths along the centrelines of elongated structures as a reinforcement learning (RL) problem, and then explore the use of deep RL (DRL) to train deep convolutional neural networks (CNNs) to learn tracking policies (Zhang et al., 2018). This removes the need for explicit appearance (observation) models and state evolution models, and additionally enables potentially richer objectives to be optimised.

We extend prior work in several ways. Firstly, we utilise a continuous action space, in contrast to prior work using DRL in biomedical imaging (Ghesu et al., 2016; Maicas et al., 2017; Liao et al., 2017; Krebs et al., 2017; Alansary et al., 2018; Zhang et al., 2018; Al and Yun, 2018; Ghesu et al., 2019), to perform subpixel tracking. Secondly, unlike Zhang et al. (2018), we address the challenge of limited training data, which is common in biomedical settings; we train our model on a simple synthetic dataset and test it on an axonal mouse cortex dataset (Bass et al., 2017), which contains many crossing elongated structures. The results of our tracker, which has to perform transductive (Pan et al., 2010) or “zero-shot” transfer to the microscopy data, can be seen in Figure 1. Despite this, our model’s performance approaches that of the current standard in the field, the Vaa3D tracker (Peng et al., 2010). By fine-tuning on the real data, which is viable with a small amount of labelled data in a mainly unlabelled dataset, we can improve performance even further. Finally, we demonstrate that the entropy of our model’s outputs—a measure of uncertainty—corresponds with how well it stays on track. Such a property is valuable in biomedical contexts, and is often lacking from many trackers, including Vaa3D’s. Our work represents another step towards learning general trackers for biomedical images, and we have open sourced our code<sup>1</sup> and data<sup>2</sup> to support further research in this direction.

---

1. [https://bitbucket.org/bicv/axon\\_tracking\\_with\\_rl/](https://bitbucket.org/bicv/axon_tracking_with_rl/)  
2. <https://www.zenodo.org/record/1182487>

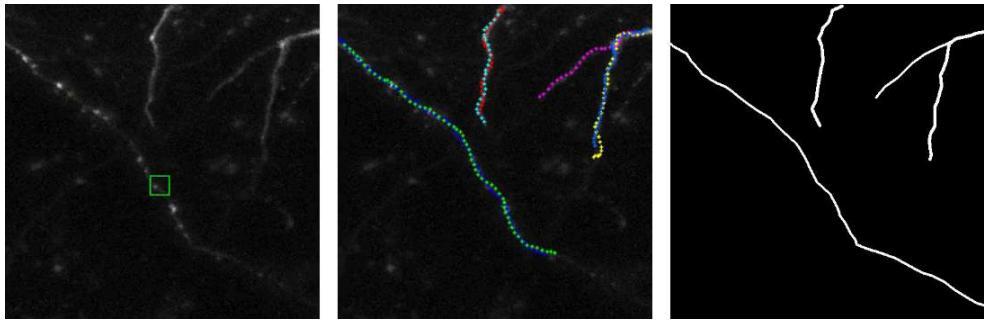


Figure 1: The operation of our learned tracker (left) on a microscopy image of several axons (Bass et al., 2017). After starting the tracker from manually labelled end-points, we can construct traces of all found axons, including branches (centre). A manual segmentation of the axons (right) is provided for comparison.

## 2. Background

### 2.1. Reinforcement Learning

RL is a branch of machine learning in which the objective is to learn to make an optimal sequence of decisions in order to maximise a reward (Sutton and Barto, 1992). In our case, the task is to trace the centreline of a biological structure from one end to another. Given the wide variety in the appearance of these structures, as well as the imaging conditions, our aim is to develop a learning agent that can be trained on the data in question, rather than having to manually tune a fixed tracking algorithm on each new dataset. We shall now explain RL more generally, as well as our specific algorithmic choices.

In RL, an agent inhabits an environment, and makes a sequence of decisions based on what it observes. At every timestep  $t$ , the agent receives the current state of the environment,  $s_t$ , and chooses an action,  $a_t$ , according to its policy  $\pi$ —a probability distribution that maps states to actions. As a result of taking an action, the agent receives a new state together with a scalar reward  $r_{t+1}$ , which gives it information about its performance. The goal of the agent is to maximise its expected return,  $\mathbb{E}[R]$ , in episodes of length  $T$ , where  $R = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$ . Here,  $\gamma \in [0, 1]$  is a discount factor which can be tuned to prioritise immediate rewards over more distant ones.

To solve our RL problem, we use an actor-critic algorithm, which combines learning a policy (actor) and value function (critic) (Sutton and Barto, 1992). Moreover, we use DRL, modelling both the policy and value functions with CNNs.<sup>3</sup> The use of CNNs in RL agents allows them to learn directly from images, with successful applications including real-world visual navigation (Arulkumaran et al., 2017).

In order to achieve *subpixel* accuracy when tracing centrelines, the agent should ideally output continuous (real-valued) actions that can be mapped to subpixel coordinates. We therefore use proximal policy optimization (PPO) (Schulman et al., 2017), a state-of-the-art actor-critic algorithm which supports continuous action policies. Its main benefit is the use of an adaptive penalty on the

3. The CNN architectures can be found in Appendix A.

amount by which the policy can change during an update, which results in more stable training than many other RL algorithms.

The original PPO algorithm used Gaussian distributions, which have infinite support ([Schulman et al., 2017](#)). In practice, we may want to constrain the action space, for example to prevent the agent from traversing many pixels in one action. Rather than penalising or thresholding large actions, which can adversely bias learning, we instead use beta distributions, which have finite support and have empirically been shown to improve the convergence of DRL algorithms on a range of problems ([Chou et al., 2017](#)).

Actor-critic algorithms make use of a learned value function to reduce the variance of the policy updates ([Schulman et al., 2016](#)). We use a particular form of variance reduction technique known as generalised advantage estimation (GAE) ([Schulman et al., 2016](#)), which introduces a small bias in return for extra variance reduction.

In actor-critic methods, the value function is only needed during training. [Pinto et al. \(2017\)](#) introduced the asymmetric actor-critic algorithm, in which the critic is given extra information that is available only during training, which can improve the learning of the true value function. Consequentially, the actor then benefits from less biased policy updates. In our setup, we achieve this by giving the critic access to the “ground truth” (the centreline, which acts as a visual representation of the reward), which allows it to learn the underlying value function more easily. The actor does not view the ground truth; and once trained, it operates directly upon on the image data, with no explicit segmentation.

## 2.2. Biomedical Imaging

There are several areas within the field of biomedical imaging that are related to our work. We will briefly cover research on thin structures within biomedical images (e.g., axons, neurons and vessels), tracking, and other uses of DRL for image analysis tasks.

### 2.2.1. THIN STRUCTURES

The analysis of elongated anatomical structures such as neurons, retinal vasculature, and muscle fibres, are important in the fields of medicine, physiology and neuroscience for diagnosis, and for the study of biological processes. These thin structures have varying properties, due to the differences in imaging techniques (microscopy, ultrasound), conditions (lighting, noise), and due to the underlying biological variability. For example, whole neurons have cell bodies, as well as dendrites and axons containing synapses which are blob-like in appearance, while retinal vasculature has a tree like structure, often with many branches ([Fraz et al., 2012](#)). Much of the prior work has focused on using segmentation algorithms for the analysis of these datasets. Segmentation of retinal vasculature has been extensively studied, and we refer the reader to [Fraz et al. \(2012\)](#) and [Kirbas and Quek \(2003\)](#) for detailed reviews of the topic. More recent works have used deep neural networks to segment thin structures in vessels ([Maninis et al., 2016](#)), muscle fibres ([Farris and Lichtwark, 2016](#); [Xie et al., 2016](#); [Cunningham et al., 2017](#)), whole neurons ([Zhou et al., 2018](#); [Li et al., 2017](#); [Liu et al., 2018](#)), and in axons ([Bass et al., 2019](#)). Of particular relevance is the work of [Li et al. \(2017\)](#), which involved both segmentation and tracking. They used a CNN for the segmentation of whole neurons, followed by a tracking algorithm to extract a graph structure from the segmentation map. While successful, this type of approach relies on an accurate segmentation for tracking to succeed.

### 2.2.2. TRACKING

Tracking in biomedical images has been used for a range of applications, including tracing elongated structures such as neurons (Meijering, 2010; Skibbe et al., 2019), vessels (Fraz et al., 2012), and muscle fibres (Farris and Lichtwark, 2016). Tracking can be used either with or without prior segmentation in situations where (i) one wants to quantify the tracked structure, (ii) the proportion of pixels representing the structure of interest is small (Helmstaedter et al., 2008) and (iii) there are branches, terminations or obscuring structures, such as blood vessels (Fraz et al., 2012). The path established by a tracker can be used to order and capture quantitative measures about morphology of entities, such as width, direction, the presence and number of branches, information that requires additional processing if segmentation is used. Moreover, tracking methods can identify branching points over the course of tracing and they can maintain the identities of branches emerging from the branching points, providing information on topology and connectivity.

### 2.2.3. DEEP REINFORCEMENT LEARNING

The combination of deep neural networks with reinforcement learning (DRL) has been successfully utilised across a range of applications in the last few years (Arulkumaran et al., 2017), including biomedical imaging (Ghesu et al., 2016; Maicas et al., 2017; Liao et al., 2017; Krebs et al., 2017; Alansary et al., 2018; Zhang et al., 2018; Al and Yun, 2018; Ghesu et al., 2019). These combine both feature learning (as opposed to utilising hand-engineered appearance models) with a general optimisation objective, formulated as a sequential decision problem in order to fit into the RL framework. Prior works have included applications to landmark detection (where the agent is similarly “embodied” in the image and must find a specific structure) (Ghesu et al., 2016; Maicas et al., 2017; Al and Yun, 2018; Ghesu et al., 2019), view planning (finding optimal 2D views in 3D images for downstream tasks) (Alansary et al., 2018), and image registration (aligning images to the same coordinate system) (Liao et al., 2017; Krebs et al., 2017). While these applications allow the agent to take any path to the solution, tracking requires the path to be as close as possible to the ground truth path of the underlying (in our case, anatomical) structure at every point. Having an underlying path allows for a denser reward signal, but also leaves less room for failure. Zhang et al. (2018) previously proposed the use of DRL for centreline tracing for blood vessels. One of the major differences is their use of a discrete action space, which limits their ability to make subvoxel traces.<sup>4</sup> In addition, they are able to train directly on hundreds of real labelled images, so do not have to address a transfer learning problem. Finally (with the exception of Al and Yun (2018) who also use an actor-critic approach but still with discrete actions), all of these prior works have been based on the deep Q-network (DQN) (Mnih et al., 2015) or variants thereof, and are hence restricted to discrete action spaces; whereas we utilise DRL with a continuous action space for biomedical imaging applications. Additionally, the output of DQNs are “Q-values” as opposed to probability distributions, where the latter allows us to directly provide uncertainty estimates via the entropy of the policy.

## 3. Subpixel Neural Tracking via Reinforcement Learning

### 3.1. Environment

We now discuss how tracing centrelines in medical images can be formulated as an RL problem. This includes the environment, the state and action spaces and the reward function.

---

4. We also formulate a different reward function to specifically account for taking subpixel movements.

**Environment:** Our environment is based on 2D greyscale images of a neuron; in RL terminology we use one image per “episode”, such that the agent receives a new neuron to track every episode. For our experiments, we use synthetic and microscopy images. In both cases the background is noisy with a low average intensity, and the neuronal structures are depicted by brighter pixels. We treat the agent as being “embodied” in the environment, which means that it is positioned within the image. The agent starts at a predefined position—one end of an axon—and moves until it finds another end. If the agent does not find another end, we terminate the episode after 200 timesteps. The environment generation is highly stochastic, but with selectable degrees of complexity.

**State space:** Rather than using the whole image as input, we provide an egocentric, multiscale view to the agent, with all views at  $11 \times 11$  px. The view comprises of one window at full resolution ( $11 \times 11$  px) and one  $21 \times 21$  px window downsampled via bilinear interpolation to  $11 \times 11$  px.<sup>5</sup> In order to give temporal context to the agent, we also provide the historical path—a view which shows the previously visited pixels around the agent’s current position. The (asymmetric) critic also receives a view containing the centreline. As the agent’s location, and hence the centre of the views, is specified with subpixel accuracy, we use bilinear interpolation to provide a correctly centred viewpoint to the agent. We zero pad all images when the view extends beyond the edge of the original image. Finally, to provide further temporal context, we concatenate all the different types of views with the corresponding views from the 3 previous timesteps (Mnih et al., 2015). The full state for the agent is visualised in Figure 2.

**Action space:** Fine structures, like axons, can be smaller than the pixel size of the image, due to discretisation in the imaging process. Furthermore, data acquisition can occur at different resolutions. To account for this, our aim is to achieve subpixel tracking. There are two components to attaining subpixel accuracy. First, rather than using discrete control, we use continuous control, with actions moving the position of the agent in 2D space. This means that the estimated position as the result of any action comes in the form of floating point coordinates that potentially (indeed, usually) are in between pixel centres. The second component relates to the reward function, discussed below.

**Reward function:** Our goal is for the agent to follow the centreline of an axon, which needs to be expressed via an appropriate reward function. The base reward is the average integral of intensity between the agent’s current and next location. To achieve subpixel accuracy, the spatial distribution of image intensities is resampled at subpixel locations, using bilinear interpolation, to calculate the reward function using the integral of intensity along straight line segments. As extra heuristics that we found to be empirically useful, we also provide a negative reward if the agent does not move, and also penalise switching directions more than once (defined as an action that is  $> 90^\circ$  from the previous action).<sup>6</sup>

### 3.2. Agent

The policy and value function of the agent are represented by separate CNNs, and are depicted in Figure 2, along with their respective inputs. The output of the actor network is a set of parameters for two independent beta distributions, from which actions (displacements in the x and y coordinates of the agent) can be sampled.<sup>7</sup> During training we sample actions, but during testing we use the means of the distributions, which results in a deterministic policy for evaluation. As the support of the beta

---

5. As our synthetic data is slightly lower resolution than our real data, we use  $1.5 \times$  the window size on the real data, downsampled to  $11 \times 11$  px.

6. Reward pseudocode can be found in Appendix B.

7. We restrict the output parameters to be greater than 1 so that the beta distributions are unimodal.

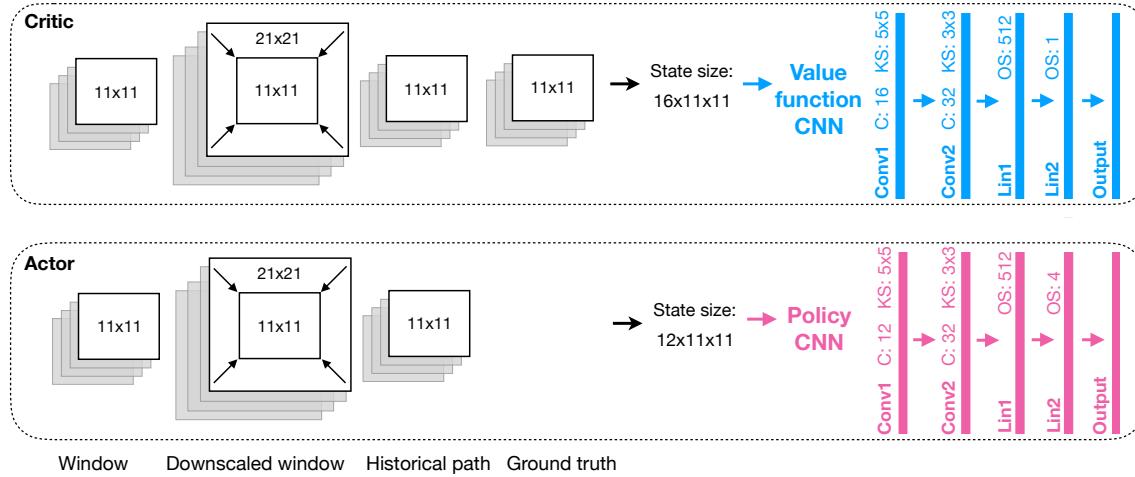


Figure 2: Inputs and architectures of the actor and the critic. All views are centered on the agent at every step, with dimensions given in pixels. The window and the downscaled window contain the pixel intensity values from the neuron image; the historical path contains only information about the agent’s previous positions and the ground truth. The state contains 4 timesteps of this information. Both CNNs contain 2 convolutional (Conv) and 2 fully-connected (Lin) layers. We also show the number of channels (C), kernel size (KS) and output size (OS).

distribution is finite, we map samples from the policy  $\in [0, 1]$  to pixel displacements  $\in [-4, 4]$  in the environment. The output of the critic network is a single value representing the value function.

Both networks are updated according to the PPO algorithm (Schulman et al., 2017) with discount  $\gamma = 0.99$ , PPO clipping value  $\epsilon = 0.2$  and GAE (Schulman et al., 2016) eligibility trace value  $\lambda = 0.95$ . After collecting a batch of 32 episodes, each network is updated 10 times within PPO’s internal loop. We use the Adam optimiser (Kingma and Ba, 2015) with a learning rate of 0.0005 and an L2 weight decay factor of 0.0003.<sup>8</sup>

## 4. Experiments

### 4.1. Datasets

We evaluated our DRL tracker on several synthetic datasets and a microscopy dataset (Figure 3). For all datasets we manually specified all start points (required for Vaa3D (Peng et al., 2010) and our DRL tracker; see subsection 4.2).

**Synthetic datasets:** Real, labelled data is often difficult to obtain (due to paucity of data, noise, etc.), and so we built a simulator to generate artificial images for training and validation. We simulated single axons by fitting polynomial splines to constrained random walks through 2D space, and adding Gaussian noise to the background. We tuned the intensity and noise settings as shown in Figure 3a to pretrain our tracker for the real data, but also trained trackers successfully on other settings,

8. Training pseudocode can be found in Appendix C.

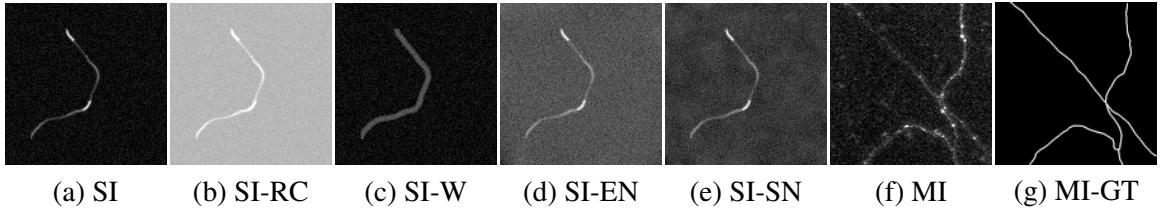


Figure 3: Examples of the 7 datasets used: synthetic (SI), synthetic reduced contrast (SI-RC), synthetic wide (SI-W), synthetic extra noise (SI-EN), synthetic structured noise (SI-SN), and microscopy images (MI). (a) is a synthetic image with the standard settings (background and axon intensities + Gaussian noise) that we use for pretraining the tracker for the MI dataset. (b), (c), (d) and (e) are synthetic images with different simulator settings. (f) was collected from a mouse somatosensory cortex using two-photon microscopy (Bass et al., 2017) with ground truth (GT) (g) labelled manually.

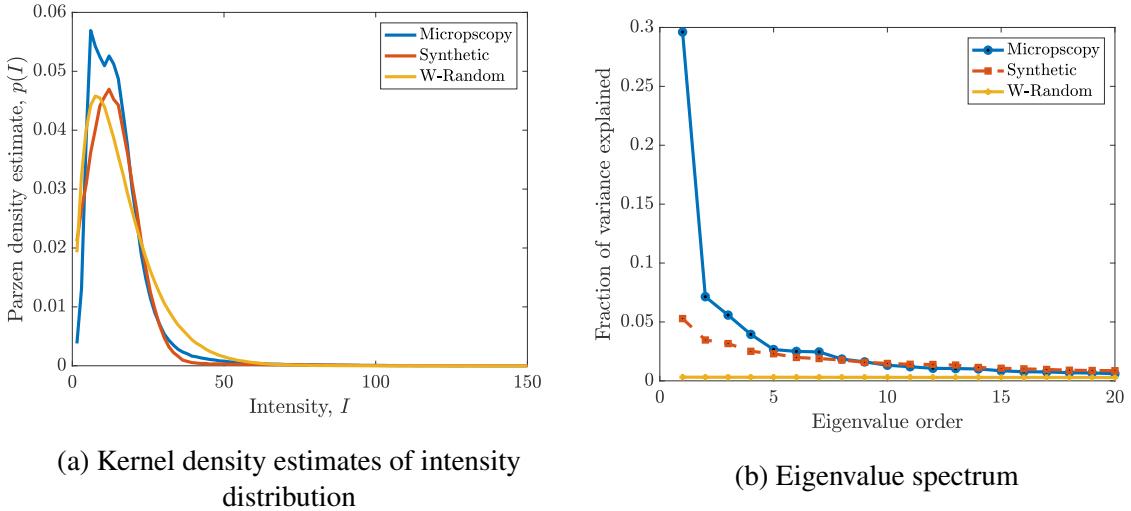


Figure 4: Quantitative comparison of synthetic and real microscopy images. (a) Kernel density estimates of pixel intensities for microscopy data vs synthetic; also shown, samples drawn from spatial white noise (W-Random) with a gamma distribution approximately matching that of real images. (b) The top 20 eigenvalue components of real data, synthetic environment, and that of white noise approximately matching the distribution of the real data. See Appendix D.1 for details.

showing that our algorithm generalises to these particular conditions: lower contrast (Figure 3b), wider structures (Figure 3c), extra noise (Figure 3d) and structured (Poisson) noise (Figure 3e).

**Microscopy dataset:** In the second stage, we tested our tracker on a mouse cortical axon dataset (Bass et al., 2017). We took a subset of 20 2D images (produced from a 3D stack), and manually

produced corresponding labels.<sup>9</sup> To generate the 2D images, we used a max projection of the 3D stack. The labels are in the form of ground truth binary images of the same size, in which the corresponding axons were segmented, and axon centerlines labelled manually using the Vaa3D software (Peng et al., 2010). We compare kernel density estimates and eigenvalue spectra of the synthetic (SI) and real images (MI) in Figure 4. As shown in Figure 4a, the SI data matches the decay in the pixel intensity values, but does not fit the low intensity pixels as well. Figure 4b demonstrates that the SI data has some level of spatial structure, but is closer to noise in comparison to the MI data.

## 4.2. Training and Testing

Training for each synthetic dataset was performed on 32,000 synthetic images.<sup>10</sup> For each synthetic dataset we validated our hyperparameters on a held-out set of 3,600 synthetic images, for a 9:1 training/validation split. We then tested our best model trained on the SI dataset on the 20 microscopy images. For our fine-tuned model, due to the lack of labelled data we used a  $k$ -fold method in which we fine-tuned the original model on 15 images and tested on the remaining 5, repeated on 4 subsets of the data.

Two measures were used to assess performance: *coverage* and *mean absolute error*. The coverage is the percentage of the axons (keypoints defined by the labels) on which the traced points were within 3px, while the error is the average perpendicular distance between axon and trace keypoints.<sup>11</sup> These metrics complement each other, as coverage quantifies how robust a tracker is (e.g., in the presence of sharp bends), while accuracy is the end goal. We compared our performance against that of trackers implemented in the Vaa3D software (Peng et al., 2010): both the default Vaa3D neuron reconstruction algorithm, which also requires start and end points as inputs, as well as the APP2 neuron tracer (Xiao and Peng, 2013), which does not require start and end points.

The coverage and error of the DRL tracker is very high and very low, respectively, on the synthetic validation sets, with little difference between the different image conditions (Table 1). The performance of the DRL tracker trained on the standard synthetic dataset (SI) drops when it is then applied to the microscopy dataset (84.10% /  $1.88 \pm 2.23$ px). This is not too far from the performance of Vaa3D, which is specialised for this kind of data (92.26% /  $0.89 \pm 0.84$ px), which is promising given that our tracker was never trained on real data, and does not incorporate any prior knowledge about microscopy data. APP2 performs the worst (81.82% /  $1.61 \pm 2.82$ px), highlighting the difficulty of performing endpoint localisation as well as tracing. We can improve coverage by fine-tuning on a very small amount of labelled data (89.08% /  $1.82 \pm 2.13$ px), suggesting that our two-phase training could be viable when a moderate amount of labelled data is available, or if an existing, hand-engineered tracker is not available at all. Average results are available in Table 2 with the per-image summary statistics available in Table 5 and the distribution of errors available in Figures 6 and 7.

As we use a stochastic policy, we can characterise the *uncertainty* of the DRL tracker by evaluating its entropy at any state.<sup>12</sup> There is a significant difference between the average en-

---

9. Note that producing centreline labels is a time-consuming process requiring an expert, restricting the amount of labelled data that we were able to procure.

10. Training on each synthetic dataset takes less than 5 hours on a GeForce 1080Ti.

11. If any tracker has an error over 11px for more than 5 consecutive timesteps we consider tracking lost and exclude these periods from our average distance error measure; we include raw results in Figure 7.

12. Note that the entropy of the beta distribution is upper-bounded by 0.

Table 1: Validation performance: coverage (%) and average error  $\pm 1$  standard deviation (px) of DRL trackers, trained on different sets of synthetic images and validated on corresponding simulator settings.

<b>SI</b>		<b>SI-RC</b>		<b>SI-W</b>		<b>SI-EN</b>		<b>SI-SN</b>	
Cov.	Error	Cov.	Error	Cov.	Error	Cov.	Error	Cov.	Error
93.37	$0.64 \pm 0.88$	94.13	$0.68 \pm 0.88$	94.62	$0.39 \pm 0.82$	96.18	$0.53 \pm 0.66$	94.72	$0.59 \pm 0.76$

Table 2: Test performance: coverage (%) and average error  $\pm 1$  standard deviation (px) of trackers, averaged over all microscopy images. DRL represents the performance of the agent trained on SI data alone, and DRL (FT) represents the performance of the DRL agent after fine-tuning on the MI data.

<b>DRL</b>		<b>DRL (FT)</b>		<b>APP2</b>		<b>Vaa3D</b>	
Cov.	Error	Cov.	Error	Cov.	Error	Cov.	Error
84.10	$1.88 \pm 2.23$	89.08	$1.82 \pm 2.13$	81.82	$1.61 \pm 2.82$	92.26	$0.89 \pm 0.84$

tropy within ( $< 3\text{px}$ ) and outside ( $\geq 3\text{px}$ ) the threshold: 0.91 on average for the DRL tracker ( $p < 0.00001$ , paired  $t$ -test), and 0.78 for the fine-tuned DRL tracker ( $p < 0.00001$ , paired  $t$ -test)<sup>13</sup>, indicating a quantitative increase in the trackers’ uncertainty if they stray from an axon. The ability to extract such a value is in contrast to traditional trackers such as Vaa3D, which typically do not quantify their uncertainty.

## 5. Conclusion

We proposed a new approach to tracking thin biological structures, such as axons and blood vessels, that is capable of producing subpixel-level traces. This first involved formulating tracking such structures as a reinforcement learning problem (Zhang et al., 2018). We then introduced a tracker based on a combination of state-of-the-art deep reinforcement learning (DRL) techniques (Schulman et al., 2017; Chou et al., 2017; Schulman et al., 2016; Pinto et al., 2017), and exposed it to an environment which simulates the physical processes of imaging. Our DRL tracker was close to the performance, with respect to coverage, of the current standard for tracing neurons (Peng et al., 2010), despite only being trained on simpler, synthetic data. Further, this is achieved without an explicit segmentation being applied as a pre-processing step.

We were able to improve coverage performance by fine-tuning the tracker on a small amount of real data, which makes our method viable for semi-supervised settings. A promising direction for future work is to incorporate synthetic data from conditional generative models during training in order to improve the realism of the synthetic data, and hence performance on real data; for example, the CapsPix2Pix model (Bass et al., 2019), from which synthetic images were utilised to improve

13. See Table 6 for the per-image breakdown.

the performance of segmentation models trained on the same microscopy dataset that we used in our work (Bass et al., 2017).

Finally, we were able to extract a quantitative measure of uncertainty, which corresponded to how well the tracker performed. While it appears that the uncertainty is well-calibrated, i.e., the values are significantly different when the agent is on- and off-track, the distribution of values also differ between images. Developing automated methods that stop tracking in test images using entropy is an interesting avenue for future work.

Our proposed method can naturally be applied to tracking elongated structures in other types biomedical images, and, furthermore, could serve as a building block for future research aimed at 3D subpixel tracking of elongated structures, for boundary trackers, and for other tasks that are currently difficult to fully automate. On the other hand, this also opens up a challenging new task for testing DRL algorithms, which are typically evaluated on video games with simple graphics (Bellemare et al., 2013) or simple control tasks with symbolic inputs (Brockman et al., 2016).

## Acknowledgments

We would like to thank Ryutaro Tanno for general feedback and Amir Alansary for help with relevant literature. We would like to also acknowledge funding from the Samsung Global Research Outreach program, the Department of Bioengineering, Imperial College London, and the Engineering and Physical Sciences Research Council [grant number EP/L016737/1].

## References

- Ludovica Acciai, Paolo Soda, and Giulio Iannello. Automated neuron tracing methods: an updated account. *Neuroinformatics*, 14(4):353–367, 2016.
- Walid Abdullah Al and Il Dong Yun. Partial policy-based reinforcement learning for anatomical landmark localization in 3d medical images. *arXiv preprint arXiv:1807.02908*, 2018.
- Amir Alansary, Loic Le Folgoc, Ghislain Vaillant, Ozan Oktay, Yuanwei Li, Wenjia Bai, Jonathan Passerat-Palmbach, Ricardo Guerrero, Konstantinos Kamnitsas, Benjamin Hou, et al. Automatic view planning with multi-scale deep reinforcement learning agents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 277–285. Springer, 2018.
- Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE SPM*, 34(6):26–38, 2017.
- Yaakov Bar-Shalom and Xiao-Rong Li. *Multitarget-multisensor tracking: principles and techniques*, volume 19. YBs London, UK:, 1995.
- Cher Bass, Pyry Helkkula, Vincenzo De Paola, Claudia Clopath, and Anil Anthony Bharath. Detection of axonal synapses in 3D two-photon images. *PLoS ONE*, 12(9):1–18, 2017. ISSN 19326203. doi: 10.1371/journal.pone.0183309.
- Cher Bass, Tianhong Dai, Benjamin Billot, Kai Arulkumaran, Antonia Creswell, Claudia Clopath, Vincenzo De Paola, and Anil Anthony Bharath. Image synthesis with a convolutional capsule generative adversarial network. In *Medical Imaging with Deep Learning*, 2019.

- Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *JAIR*, 47:253–279, 2013.
- Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.
- Po-Wei Chou, Daniel Maturana, and Sebastian Scherer. Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution. In *ICML*, pages 834–843, 2017.
- Ryan Cunningham, Peter Harding, and Ian Loram. Deep residual networks for quantification of muscle fiber orientation and curvature from ultrasound images. In *Annual Conference on Medical Image Understanding and Analysis*, pages 63–73. Springer, 2017.
- Dominic James Farris and Glen A Lichtwark. Ultratrack: Software for semi-automated tracking of muscle fascicles in sequences of b-mode ultrasound images. *Computer methods and programs in biomedicine*, 128:111–118, 2016.
- Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. Blood vessel segmentation methodologies in retinal images—a survey. *Computer Methods and Programs in Biomedicine*, 108(1):407–433, 2012.
- Florin C Ghesu, Bogdan Georgescu, Tommaso Mansi, Dominik Neumann, Joachim Horngger, and Dorin Comaniciu. An artificial agent for anatomical landmark detection in medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 229–237. Springer, 2016.
- Florin-Cristian Ghesu, Bogdan Georgescu, Yefeng Zheng, Sasa Grbic, Andreas Maier, Joachim Horngger, and Dorin Comaniciu. Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):176–189, 2019.
- Hayit Greenspan, Bram Van Ginneken, and Ronald M. Summers. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Trans. Medical Imaging*, 35(5):1153–1159, 2016. ISSN 1558254X. doi: 10.1109/TMI.2016.2553401.
- Moritz Helmstaedter, Kevin L Briggman, and Winfried Denk. 3d structural imaging of the brain with photons and electrons. *Current opinion in neurobiology*, 18(6):633–641, 2008.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- Cemil Kirbas and Francis KH Quek. Vessel extraction techniques and algorithms: a survey. In *Proc. Bioinformatics and Bioengineering (BIBE)*, pages 238–245. IEEE, 2003.
- Julian Krebs, Tommaso Mansi, Hervé Delingette, Li Zhang, Florin C Ghesu, Shun Miao, Andreas K Maier, Nicholas Ayache, Rui Liao, and Ali Kamen. Robust non-rigid registration through agent-based action learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 344–352. Springer, 2017.

Rahul Prasanna Kumar, Fritz Albregtsen, Martin Reimers, Bjørn Edwin, Thomas Langø, and Ole Jakob Elle. Blood vessel segmentation and centerline tracking using local structure analysis. In *6th European conference of the international federation for medical and biological engineering*, pages 122–125. Springer, 2015.

Rongjian Li, Tao Zeng, Hanchuan Peng, and Shuiwang Ji. Deep learning segmentation of optical microscopy images improves 3-d neuron reconstruction. *IEEE transactions on medical imaging*, 36(7):1533–1541, 2017.

Rui Liao, Shun Miao, Pierre de Tournemire, Sasa Grbic, Ali Kamen, Tommaso Mansi, and Dorin Comaniciu. An artificial agent for robust image registration. In *AAAI Conference on Artificial Intelligence*, 2017.

Min Liu, Huiqiong Luo, Yinghui Tan, Xueping Wang, and Weixun Chen. Improved v-net based image segmentation for 3d neuron reconstruction. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 443–448. IEEE, 2018.

Gabriel Maicas, Gustavo Carneiro, Andrew P Bradley, Jacinto C Nascimento, and Ian Reid. Deep reinforcement learning for active breast lesion detection from dce-mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 665–673. Springer, 2017.

Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Deep retinal image understanding. In *International conference on medical image computing and computer-assisted intervention*, pages 140–148. Springer, 2016.

Erik Meijering. Neuron tracing in perspective. *Cytometry Part A*, 77(7):693–704, 2010.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei a Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. ISSN 0028-0836. doi: 10.1038/nature14236.

Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Hanchuan Peng, Zongcai Ruan, Fuhui Long, Julie H. Simpson, and Eugene Myers. V3d enables real-time 3d visualization and quantitative analysis of large-scale biological image data sets. *Nature Biotechnology*, 28(4):348–353, 2010. doi: 10.1038/nbt.1612.

Hanchuan Peng, Michael Hawrylycz, Jane Roskams, Sean Hill, Nelson Spruston, Erik Meijering, and Giorgio A. Ascoli. BigNeuron: Large-Scale 3D Neuron Reconstruction from Optical Microscopy Images, 2015. ISSN 10974199.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. *arXiv preprint arXiv:1710.06542*, 2017.

- Philippe Poulin, Marc-Alexandre Cote, Jean-Christophe Houde, Laurent Petit, Peter F Neher, Klaus H Maier-Hein, Hugo Larochelle, and Maxime Descoteaux. Learn to track: Deep learning for tractography. In *MICCAI*, pages 540–547. Springer, 2017.
- John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation. *arXiv*, pages 1–9, 2016.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Henrik Skibbe, Marco Reisert, Ken Nakae, Akiya Watakabe, Junichi Hata, Hiroaki Mizukami, Hideyuki Okano, Tetsuo Yamamori, and Shin Ishii. Pat—probabilistic axon tracking for densely labeled neurons in large 3-d micrographs. *IEEE transactions on medical imaging*, 38(1):69–78, 2019.
- R.S. Sutton and A.G. Barto. *Reinforcement Learning*. Springer US, Boston, MA, sep 1992. ISBN 978-1-4613-6608-9. doi: 10.1007/978-1-4615-3618-5.
- Harry L Van Trees and Kristine L Bell. Bayesian bounds for parameter estimation and nonlinear filtering/tracking. *AMC*, 10:12, 2007.
- Hang Xiao and Hanchuan Peng. App2: automatic tracing of 3d neuron morphology based on hierarchical pruning of a gray-weighted image distance-tree. *Bioinformatics*, 29(11):1448–1454, 2013.
- Yuanpu Xie, Zizhao Zhang, Manish Sapkota, and Lin Yang. Spatial clockwork recurrent neural network for muscle perimysium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 185–193. Springer, 2016.
- Pengyue Zhang, Fusheng Wang, and Yefeng Zheng. Deep reinforcement learning for vessel centerline tracing in multi-modality 3d volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 755–763. Springer, 2018.
- Zhi Zhou, Hsien-Chi Kuo, Hanchuan Peng, and Fuhui Long. Deepneuron: an open deep learning toolbox for neuron tracing. *Brain informatics*, 5(2):3, 2018.

## Appendix A. Neural Network Architectures

Tables 3 and 4 contain the network architectures for the policy and value networks, respectively. We use ReLU nonlinearities between all hidden layers.

Table 3: Actor/policy network. The output of the network is two pairs of  $\alpha$  and  $\beta$  parameters for two beta distributions corresponding to displacements in the x and y coordinates. We use a softplus nonlinearity and add 1 to the outputs of the final layer to ensure that these parameters are  $> 1$ ; this is done so that the beta distributions are always unimodal.

Layer type	Input channels	Input size	Kernel size	Stride	Padding	Output channels	Output size
Convolution	12	$11 \times 11$	$5 \times 5$	1	2	32	$11 \times 11$
Convolution	32	$11 \times 11$	$3 \times 3$	1	1	32	$13 \times 13$
Linear	-	5408	-	-	-	-	512
Linear	-	512	-	-	-	-	4

Table 4: Critic/value network. The output of the network is a single real number representing the state's value function.

Layer type	Input channels	Input size	Kernel size	Stride	Padding	Output channels	Output size
Convolution	16	$11 \times 11$	$5 \times 5$	1	2	32	$11 \times 11$
Convolution	32	$11 \times 11$	$3 \times 3$	1	1	32	$13 \times 13$
Linear	-	5408	-	-	-	-	512
Linear	-	512	-	-	-	-	1

## Appendix B. Reward Function

---

**Algorithm 1:** DRL Axon Tracking Algorithms - Reward Calculation
 

---

```

1: Input: current agent position  $p_t$ , current action  $a_t$ , previous action  $a_{t-1}$ , ground truth centreline
   image  $I_G$ , counter  $k$ 
2: Output: reward  $r_{t+1}$ 
3: Next agent position  $p_{t+1} \leftarrow p_t + a_t$ 
4: Distance travelled  $d \leftarrow \|p_t - p_{t-1}\|$ 
5: if  $d == 0$  then
6:    $r_{t+1} = -1$ 
7: else
8:   Sample 100 points  $S$  between  $p_t$  and  $p_{t+1}$ 
9:   Reward  $r_{t+1} \leftarrow \sum_{s=1}^{|S|} I_G(x_s, y_s)$ 
10:  if angle between  $a_{t-1}$  and  $a_t > 90^\circ$  then
11:     $k \leftarrow k + 1$ 
12:  end if
13:  if  $k \pmod{2} == 1$  then
14:     $r_{t+1} \leftarrow -r_{t+1}$ 
15:  end if
16: end if
return  $r_t$ 
  
```

---

## Appendix C. Training Algorithm

---

### Algorithm 2: DRL Axon Tracking Algorithms - Training

---

```

1: Set the batch size  $M$  and the max length of each episode  $N$ 
2: Initialize the actor-network  $A(s_a|\theta^a)$ 
3: Initialize the critic-network  $C(s_c|\theta^c)$ 
4: repeat
5:   for  $i = 1$  to  $M$  do
6:     Initialize the position of agent  $p_0$ 
7:     Initialize the counter  $k \leftarrow 0$ 
8:     Initialize the initial unit direction vector  $v_0$ 
9:     Initialize the initial state for actor-network  $s_a$ 
10:    Initialize the initial state for critic-network  $s_c$ 
11:    for  $t = 1$  to  $N$  do
12:      Select the action  $a_t = A(s_a|\theta^a)$ ;
13:      Input the action  $a_t$  into the simulator and get  $s'_a, s'_c, r_t$ , terminal;
14:      Store the transition  $(s_a, s_c, r_t, \text{terminal})$ ;
15:      if terminal then
16:        break;
17:      end if
18:       $s_a \leftarrow s'_a, s_c \leftarrow s'_c$ ;
19:    end for
20:  end for
21:  Update the actor-network  $A(s_a|\theta^a)$  by using PPO
22:  Update the critic-network  $C(s_c|\theta^c)$  by using PPO
23: until  $A(s_a|\theta^a)$  and  $C(s_c|\theta^c)$  are converged

```

---

## Appendix D. Further Experimental Results

### D.1. Plausibility of Synthetic Environment

The use of synthetic environments for training DRL agents is common in robotics and autonomous driving, but less so for image interpretation. In any use of simulation environments, the size of the “reality gap” is a key factor in determining the ability to transfer the agent into a real environment or task. Here, we describe the experiments to compare the synthetic environment with real microscopy data.

We drew 20 images from the synthetic environments, and compared pixel intensity statistics and second-order spatial statistics through eigen spectral analysis of the covariance matrices estimated from randomly selected image patches. We also applied this analysis to white noise spatial fields, and to 20 real microscopy images.

Firstly, from each of the 20 synthetic and 20 microscopy images analysed, we randomly selected 200  $25 \times 25$  pixel spatial patches, yielding 4,000 patches. We then applied a kernel density estimate to this data, using a Gaussian kernel of bandwidth 0.5 to produce estimates of image intensity in Figure 4a.

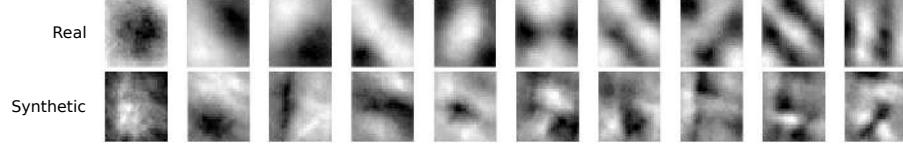


Figure 5: Eigenimages produced by finding the most important 10 eigenvalues for real image patches and for synthetic patches. Note the similarity of the first 5 eigenimages in the top and bottom rows, though there is a polarity change visible in many patches, and clear differences for eigenimages 6-10.

Once we had the histogram of pixel intensities for real images, we fit a gamma distribution to the intensity data, obtaining tight confidence intervals for the parameters (shape parameter  $1.934 \pm 0.003$  with 95% confidence, and scale parameter of  $7.699 \pm 0.014$  with 95% confidence). We then used these parameters to synthesise gamma-distributed white noise fields.

We also estimated the covariance matrix for each of the 3 sets of 4,000 image patches. The eigenvalue spectra (eigenvalue normalised by sum of absolute eigenvalues) for the top 20 components are shown in Figure 4b. The associated top 10 eigenimages corresponding to the microscopy and synthetic environments are presented in Figure 5. The eigenimages corresponding to white noise are essentially noise fields, and are not shown.

## D.2. Error analysis during tracking

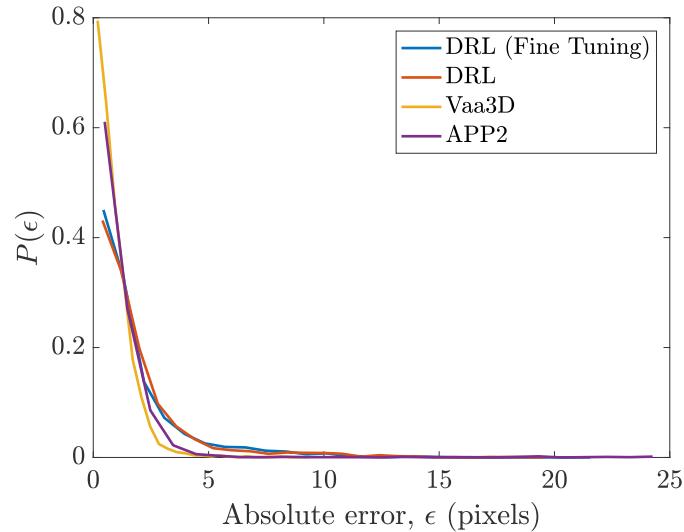


Figure 6: Distribution of errors ( $p(\text{error})$  vs.  $\text{error} (\text{px})$ ) accumulated over 3,400 track locations in all 20 images.

Table 5: Coverage (%) and average error  $\pm 1$  standard deviation (px) of trackers on individual microscopy images.

ID	DRL		DRL (fine-tuned)		APP2		Vaa3D	
	Cov.	Error	Cov.	Error	Cov.	Error	Cov.	Error
1	87.24	$1.49 \pm 1.60$	83.73	$1.94 \pm 2.50$	98.88	$1.58 \pm 2.10$	95.93	$1.00 \pm 0.91$
2	86.01	$1.41 \pm 1.40$	89.66	$1.62 \pm 1.80$	79.68	$1.47 \pm 1.70$	91.97	$1.11 \pm 1.40$
3	94.24	$1.90 \pm 2.30$	94.96	$1.78 \pm 2.60$	99.28	$0.93 \pm 1.10$	98.32	$0.79 \pm 0.60$
4	95.75	$1.22 \pm 0.94$	92.99	$1.10 \pm 1.30$	94.27	$0.75 \pm 0.96$	99.79	$0.88 \pm 1.20$
5	94.68	$1.43 \pm 1.10$	85.69	$2.58 \pm 2.60$	91.41	$1.50 \pm 2.30$	95.71	$0.72 \pm 0.80$
6	78.16	$3.78 \pm 3.50$	92.29	$1.78 \pm 2.00$	92.51	$1.47 \pm 2.60$	91.22	$0.90 \pm 0.61$
7	92.27	$1.75 \pm 1.20$	92.05	$1.45 \pm 1.40$	97.73	$1.15 \pm 1.30$	95.91	$0.89 \pm 0.79$
8	75.72	$2.06 \pm 1.60$	83.39	$2.06 \pm 1.60$	81.47	$1.13 \pm 1.00$	84.35	$1.11 \pm 0.80$
9	49.26	$3.79 \pm 4.20$	92.65	$3.50 \pm 3.10$	35.54	$3.98 \pm 6.00$	94.61	$1.49 \pm 1.10$
10	94.18	$1.50 \pm 0.49$	94.63	$1.36 \pm 0.41$	56.82	$2.13 \pm 2.1$	97.54	$1.05 \pm 0.77$
11	92.22	$4.49 \pm 4.50$	95.56	$2.36 \pm 2.60$	92.96	$2.10 \pm 3.80$	98.52	$0.73 \pm 0.48$
12	72.43	$2.06 \pm 2.60$	76.47	$1.46 \pm 1.40$	79.82	$1.13 \pm 1.70$	90.54	$1.01 \pm 0.82$
13	94.40	$1.24 \pm 0.96$	96.86	$1.37 \pm 1.50$	85.93	$3.20 \pm 4.70$	96.72	$0.82 \pm 0.68$
14	88.73	$3.06 \pm 2.70$	95.77	$3.69 \pm 3.30$	100.00	$1.22 \pm 1.30$	95.21	$1.06 \pm 0.88$
15	88.12	$1.41 \pm 1.40$	86.42	$1.97 \pm 2.40$	73.92	$1.76 \pm 2.70$	68.06	$0.74 \pm 0.57$
16	96.95	$1.13 \pm 1.10$	97.97	$1.25 \pm 2.60$	98.99	$0.61 \pm 0.45$	99.32	$0.68 \pm 0.55$
17	70.14	$1.59 \pm 1.80$	73.44	$1.58 \pm 1.60$	86.06	$2.54 \pm 4.80$	89.46	$0.81 \pm 0.67$
18	66.18	$2.10 \pm 2.60$	70.70	$1.57 \pm 1.80$	48.77	$0.62 \pm 0.57$	67.94	$0.64 \pm 0.50$
19	98.38	$0.82 \pm 0.75$	99.08	$1.05 \pm 0.79$	99.31	$0.60 \pm 0.58$	99.77	$0.55 \pm 0.42$
20	66.85	$1.64 \pm 1.90$	87.23	$1.93 \pm 2.10$	43.00	$0.56 \pm 0.53$	94.40	$0.71 \pm 0.64$
Total	84.10	$1.88 \pm 2.23$	89.08	$1.82 \pm 2.13$	81.82	$1.61 \pm 2.82$	92.26	$0.89 \pm 0.84$

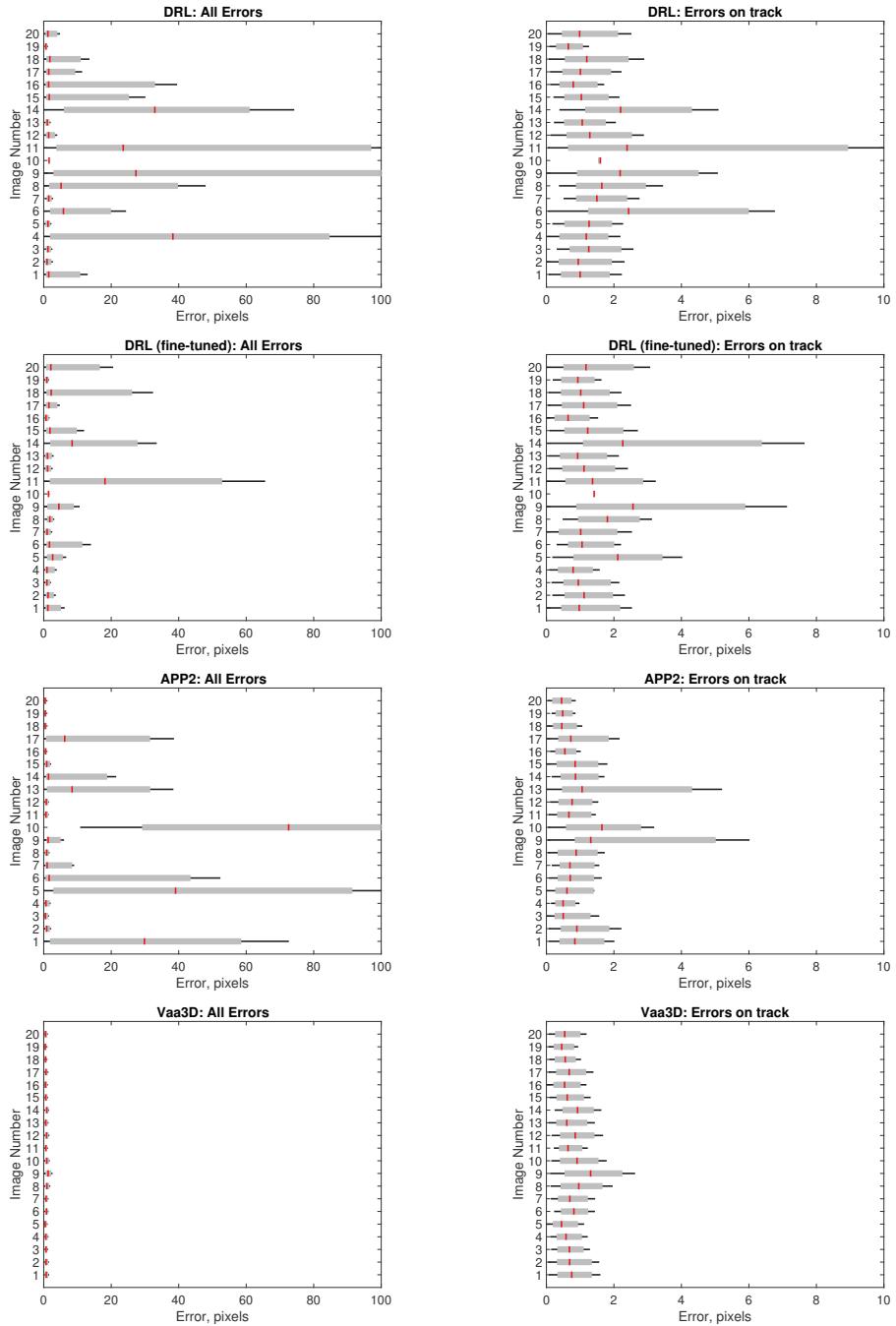


Figure 7: Error distribution (px) of trackers on individual microscopy images, without filtering criterion (left; error range of 0-100px) and with (right; error range of 0-10px). Whiskers set to  $0.25 \times$  inter-quartile range, for purposes of visibility and comparison.

Table 6: Average entropy  $\pm$  1 standard deviation within and outside the threshold ( $< 3\text{px}/\geq 3\text{px}$ ). The entropy significantly increases as the tracker passes the threshold for both the DRL tracker ( $p < 0.00001$ ) and the DRL tracker after fine-tuning (FT) ( $p < 0.00001$ ). p-values were calculated using a two-tailed paired  $t$ -test on all points within all images.

<b>ID</b>	<b>DRL</b>		<b>DRL (FT)</b>	
	Within	Outside	Within	Outside
1	$-4.63 \pm 0.83$	$-3.59 \pm 0.99$	$-3.39 \pm 0.67$	$-2.27 \pm 1.13$
2	$-4.17 \pm 0.79$	$-3.43 \pm 0.92$	$-2.83 \pm 0.56$	$-2.06 \pm 0.84$
3	$-4.16 \pm 0.64$	$-3.89 \pm 0.65$	$-2.97 \pm 0.55$	$-2.97 \pm 0.62$
4	$-4.65 \pm 0.93$	$-2.41 \pm 0.80$	$-2.79 \pm 0.62$	$-1.08 \pm 0.91$
5	$-3.96 \pm 0.87$	$-4.09 \pm 1.11$	$-2.63 \pm 0.77$	$-1.21 \pm 1.09$
6	$-3.43 \pm 0.93$	$-2.31 \pm 0.70$	$-2.13 \pm 0.68$	$-0.95 \pm 0.58$
7	$-4.34 \pm 0.82$	$-3.75 \pm 0.84$	$-3.18 \pm 0.60$	$-2.84 \pm 0.51$
8	$-3.86 \pm 0.83$	$-2.60 \pm 0.80$	$-2.75 \pm 0.51$	$-2.52 \pm 0.50$
9	$-2.85 \pm 0.75$	$-2.30 \pm 0.85$	$-2.36 \pm 0.60$	$-1.77 \pm 0.84$
10	$-3.60 \pm 0.32$	$-4.63 \pm 1.30$	$-3.35 \pm 0.28$	$-3.80 \pm 0.0$
11	$-4.23 \pm 0.90$	$-2.58 \pm 0.71$	$-3.34 \pm 0.63$	$-2.03 \pm 0.69$
12	$-3.65 \pm 0.87$	$-2.32 \pm 0.70$	$-2.94 \pm 0.62$	$-2.35 \pm 0.46$
13	$-4.21 \pm 0.94$	$-3.84 \pm 0.88$	$-3.41 \pm 0.56$	$-2.75 \pm 0.58$
14	$-4.26 \pm 0.82$	$-2.99 \pm 0.66$	$-3.43 \pm 0.53$	$-2.59 \pm 0.74$
15	$-4.51 \pm 0.79$	$-2.95 \pm 0.89$	$-3.55 \pm 0.45$	$-2.77 \pm 0.55$
16	$-3.28 \pm 1.02$	$-2.52 \pm 0.54$	$-2.63 \pm 0.53$	$-1.97 \pm 0.68$
17	$-4.16 \pm 0.88$	$-2.25 \pm 0.72$	$-3.23 \pm 0.55$	$-2.21 \pm 0.78$
18	$-3.58 \pm 1.20$	$-2.25 \pm 0.67$	$-2.75 \pm 0.62$	$-1.38 \pm 0.75$
19	$-4.77 \pm 0.68$	$-4.54 \pm 1.17$	$-3.33 \pm 0.44$	$-2.91 \pm 0.24$
20	$-3.54 \pm 1.30$	$-2.39 \pm 0.84$	$-2.75 \pm 0.78$	$-1.85 \pm 0.73$
Total	$-3.99 \pm 0.49$	$-3.08 \pm 0.79$	$-2.99 \pm 0.38$	$-2.21 \pm 0.70$

# Stain-Transforming Cycle-Consistent Generative Adversarial Networks for Improved Segmentation of Renal Histopathology

**Thomas de Bel<sup>1</sup>**

T.DEBEL@RUMC.NL

<sup>1</sup> Department of Pathology, Radboud University Medical Center, Nijmegen, the Netherlands

**Meyke Hermsen<sup>1</sup>**

MEYKE.HERMSEN@RUMC.NL

**Jesper Kers<sup>2</sup>**

J.KERS@AMC.UVA.NL

<sup>2</sup> Department of Pathology, Amsterdam University Medical Center, the Netherlands

**Jeroen van der Laak<sup>1</sup>**

JEROEN.VANDERLAAK@RUMC.NL

**Geert Litjens<sup>1</sup>**

GEERT.LITJENS@RUMC.NL

## Abstract

The performance of deep learning applications in digital histopathology can deteriorate significantly due to staining variations across centers. We employ cycle-consistent generative adversarial networks (cycleGANs) for unpaired image-to-image translation, facilitating between-center stain transformation. We find that modifications to the original cycleGAN architecture make it more suitable for stain transformation, creating artificially stained images of high quality. Specifically, changing the generator model to a smaller U-net-like architecture, adding an identity loss term, increasing the batch size and the learning all led to improved training stability and performance. Furthermore, we propose a method for dealing with tiling artifacts when applying the network on whole slide images (WSIs). We apply our stain transformation method on two datasets of PAS-stained (Periodic Acid-Schiff) renal tissue sections from different centers. We show that stain transformation is beneficial to the performance of cross-center segmentation, raising the Dice coefficient from 0.36 to 0.85 and from 0.45 to 0.73 on the two datasets.

**Keywords:** Deep learning, generative adversarial networks, medical imaging, stain transformation

## 1. Introduction

Staining of tissue is a central part of histopathology, highlighting tissue structures crucial for diagnosis. The staining process is subject to high variability. Differences can be introduced, among others, by variations in staining protocols between pathology centers and differing whole-slide scanners. Large variety in stainings has been shown to dramatically affect the performance of deep learning image analysis (Ciompi et al., 2017). To a large extent, dissimilarities between centers can be accounted for by training with color/stain augmentations or using data from multiple centers (Tellez et al., 2018). However, it is uncertain whether data augmentations are able to capture all variations that occur ‘in the wild’ due to the linear nature of many color and stain augmentations. This may be an oversimplification of the variability that occurs in real-world tissue stainings.

Once deep learning algorithms are introduced in the workflow of the pathologist, they need to achieve reliable performance, regardless of the center they are deployed. Using only augmentations, the robustness of a network is unchangeable at test time. Even if algorithms would be optimized or tuned for a specific center, newly introduced staining protocols or whole-slide scanners could result in algorithm performance degradation. This could only be resolved by retraining the algorithm for

such modifications, which is cumbersome and time-consuming. An alternate strategy is to normalize whole-slide images to mimic the data that a network was trained on, alleviating the need for algorithm re-training.

Most previous work on stain normalization focuses on hand-engineered methods. These methods are typically tuned for a specific stain, for example haematoxylin and eosin (H&E) (Bejnordi et al., 2016; Khan et al., 2014). Recent approaches have used cycle-consistent generative adversarial networks (cycleGANs), and have shown the effectiveness of this architecture when used for stain transformation (Gadermayr et al., 2018; Shaban et al., 2018).

In a GAN setup, a discriminator network  $\mathbf{D}$  is used to adversarially learn a generator  $\mathbf{G}$  a domain mapping  $G : X \rightarrow Y$ . CycleGANs add to this by introducing an inverse mapping  $F : Y \rightarrow X$  and enforce  $F(G(X)) \approx X$ , to retain structural information while transferring domains. CycleGANs are trained in an unsupervised and unpaired manner and are ‘stain-agnostic’, i.e. they can be applied for normalization of any stain.

CycleGANs may prove to be a solution to real-world stain variations, by transforming whole-slide images to the exact same stain. This would resolve the need for in-network stain robustness for algorithms that perform, for instance, cancer detection. A deployment setup can be imagined where cycleGANs are trained ‘just-in-time’ for new stain variations, after which a chain of other networks is executed to perform a variety of tasks (e.g. detection, segmentation, grading) without re-training.

**Contributions:** In this paper we make several contributions to existing work applying cycleGANs to stain transformation:

- We show that the original cycleGAN architecture benefits from optimization for stain transformation. We introduce several changes to the generator part of the cycleGAN architecture, reducing the amount of parameters of the transformation network. We tune the learning rate, batch size, and add an extra identity loss term that stabilizes training. We show that these changes result in improved stain-transferred images.
- We introduce a novel method for applying cycleGANs to whole-slide images for stain transfer. In short, this method works in a fully convolutional fashion at inference time. By sliding through the whole slide image, using weighted merging of overlapping adjacent tiles to remove tiling artifacts that would occur in regular patch-by-patch application.
- We demonstrate the effectiveness of stain transformation for cross-center tissue segmentation with convolutional neural networks. A segmentation network is trained on a dataset from one center and then applied to a test dataset from a different center. We compare the performance on test dataset with and without stain transformation. We also train the segmentation network with and without augmentations in an attempt to assess how well augmentations capture stain variation.

## 2. Experiments

### 2.1. Quantitative Analysis

Central to our method will be the performance of a segmentation network trained on data from the Radboud University Medical Centre (RUMC), Nijmegen, the Netherlands, and tested on data from the Academic Medical Center (AMC), Amsterdam. We refer to the section ‘3.1’ for a detailed

description of the data. The segmentation network is trained twice: once with and once without extensive color and spatial augmentations. We apply both versions of the segmentation network on the AMC dataset to assess the effectiveness of the augmentations. On top of this, we add our cycleGAN stain normalization. The stain normalization is trained with both datasets, to learn the transformation from both RUMC to AMC and vice versa. Again, we apply both the augmented and non-augmented versions of the segmentation network on the AMC dataset, this time after performing stain normalization. This allows us to compare augmenting vs. not augmenting in conjunction with normalization vs. no normalization. Additionally, we train two segmentation networks on the AMC dataset and perform the same four experiments on the RUMC data.

Because we introduce several alterations to the original cycleGAN architecture, we compare the performance of our algorithm with the baseline cycleGAN architecture that was used in the original paper ([Zhu et al., 2017](#)).

## 2.2. Models

**Segmentation network:** For the segmentation network we used a standard U-net, based on [Ronneberger et al. \(2015\)](#). During training, patches were sampled at roughly  $1.0 \mu\text{m}$  per pixel, with a patch size of 412,412. Apart from standard flipping and rotating, extensive color augmentations (e.g. brightness, contrast, HSV color shift) were used in an attempt to enhance the robustness to unseen stains. Figure 1 shows an example patch with the color augmentations that were performed. During training of the segmentation network, the augmentations were randomly combined to induce even more variation.

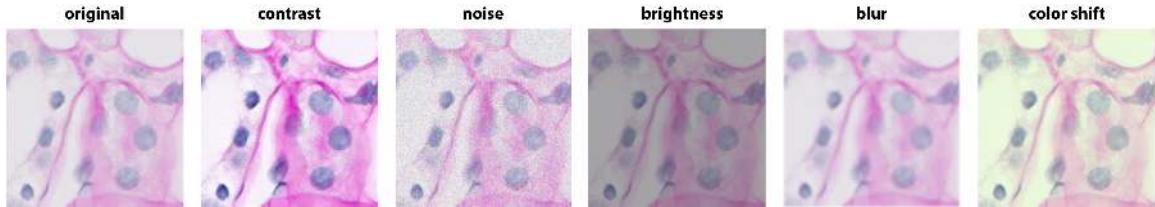


Figure 1: A sample of the color augmentations that were performed during training of the segmentation network.

**Stain transformation network:** The baseline cycleGAN-architecture is optimized using the cycle-consistency loss. The baseline adopts the generator  $\mathbf{G}$ ) of [Johnson et al. \(2016\)](#) for its good results on neural style transfer. Similar  $\mathbf{G}$  networks were also utilized in recent cycleGAN stain transformation approaches ([Shaban et al., 2018; Rivenson et al., 2018](#)). For our discriminator  $\mathbf{D}$ , we used the same 70x70 PatchGAN model as the baseline ([Isola et al., 2017](#)).

Our first modification to the baseline approach is increasing the batch size to 6 patches and increasing the learning rate to 0.008. As also mentioned in [Brock et al. \(2018\)](#), we hypothesize that more modi are covered in one batch, stabilizing training and reducing the probability of introducing hallucination artifacts ([Cohen et al., 2018](#)). To further stabilize training, we add an extra identity loss to the optimization process:

$$L_{identity}(G, F) = E_x[||G(x) - x||_1] + E_y[||F(y) - y||_1], \quad (1)$$

where  $\mathbf{G}$  and  $\mathbf{F}$  are the generators of both transformation directions. This loss forces the generators to perform an identity mapping of the input. As this loss discourages the generator to learn the stain transformation, the weight of this loss is gradually decreased to zero in the first 20 epochs. We found that adding this loss stabilizes training by forcing the network to initially look at 'simple' solutions close to the identity function. This prevents divergence to poor local optima in initial phases of the training process. An example of bad convergence is shown in the results section in Figure 6 (d).

We also modify both  $\mathbf{G}$  and  $\mathbf{F}$  to follow a U-net-like structure, using ResNet blocks and skip-connections between the encoder and decoder (He et al., 2016). We change the transposed convolutions in the decoder part to nearest neighbours up-sampling layers based on Odena et al. (2016). The width of the first layer starts at 32. The amount of filters is increased by a factor of two after each max-pooling layer and decreased by the same factor after each up-sampling layer. We use a small generator network with U-net depth of three, i.e. three max-pooling layers and up-sampling layers. We also experimented with further reducing the amount of parameters, by lowering the depth of the network to two and one. Both  $\mathbf{G}$  and  $\mathbf{D}$ , with each convolution, use leaky ReLU's and instance normalization, which has shown to work well for style transfer (Ulyanov et al., 2016). For cycle-consistency loss we used the  $L1$ -norm,  $\mathbf{D}$  was optimized with the mean squared error loss. We trained the networks with patches of size  $256 \times 256$ .

**Patch sampling:** We used tissue masks for sampling from the whole slide images (WSIs) during training of the cycleGAN. The masks were generated by using adaptive thresholding, with a window size of 11. During training, we uniformly sampled patches on-the-fly from the tissue based on the mask. This leads to a high variability where no two patches are exactly the same. Figure 2 shows the use of these masks with randomly sampled seed-points to generate patches.

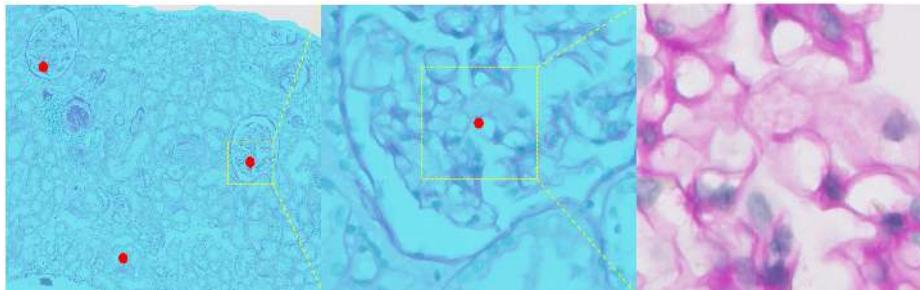


Figure 2: We generate patches on the fly from seed points samples from the mask, overlaid on the image. Taking the the seed point as central pixel, we generate a patch (of  $256 \times 256$  in our case). These seed-points and patches are generated on-the-fly during training.

### 3. WSI inference technique

As WSI images are too large to fit directly on a GPU, we perform inference tile-by-tile to obtain the stain transformed whole-slide images. This introduces artifacts between adjacent tiles in the transformed WSI due to instance normalization relying on tile statistics. One option would be to use the running mean and average values obtained during training. This would reduce the quality

of the transformation, as the individual WSIs differ in their color intensities due to stain variations within a dataset. We propose a tile-wise inference method that eliminates the tiling artifacts. First, we increase the input size of our cycleGAN to 2048 pixels during inference. This will reduce the variation in instance normalization tile statics that occur when using small patches. We subsequently crop the network output to 1024 to get rid of the border artifacts introduced by zero-padding. Second, we take overlapping tiles by only shifting 512 pixels to our next tile. These tiles will largely have the same normalization statistics due to the small shift. This is visualized in Figure 3 (a). Last, we weight the pixels in the tiles based on their distance from the center pixel of the tile, to create a smooth transition between overlapping tiles. Here, the weight for a single pixel is based on the following formula:

$$w = \min(|x - x_{cp}|, |y - y_{cp}|), \quad (2)$$

where  $cp$  stands for the center pixel. The weight map this creates for each tile is visualized in Figure 3 (b). Finally, due to overlap in both the  $x$  and  $y$  direction, there are four weighted values per pixel. We sum the weighted pixel values and normalize by the sum of the weights to create the final result. The effect of these tiling strategies can be seen in Figure 4.

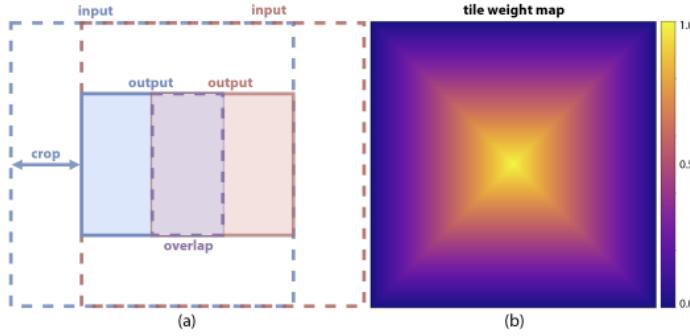


Figure 3: Schematic overview of the WSI inference strategy. (a) Shows the sliding window technique. Inference is performed on a large input, after which half of the image is cropped to remove zero-padding artifacts. We shift the window by half of the output (cropped) size, creating an overlap between tiles. (b) Visualizes the tile weight map.

### 3.1. Evaluation

**Data:** We utilize two datasets with periodic acid-Schiff (PAS) stains. The first dataset consists of forty biopsies originating from RUMC. The tissue slides were digitized using the 3D Histech’s Panoramic 250 Flash II scanner. The second dataset consists of twenty-four biopsies, stained at the AMC. The slides were scanned with the Philips IntelliSite Ultra Fast Scanner. All slides were scanned at roughly  $0.25 \mu\text{m}$  per pixel. Figure 5 shows an example of RUMC and AMC PAS-stained tissue. We included seven structure classes in our segmentation task: glomeruli, empty glomeruli, sclerotic glomeruli, distal tubuli, proximal tubuli, atrophic tubuli and arteries. All pixels within the regions of interest that did not belong to any category, were put in an eighth background structure class. Ten slides of the AMC dataset and forty slides of the RUMC dataset were annotated

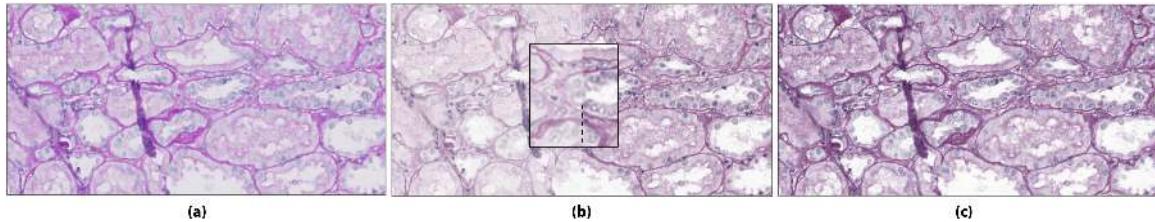


Figure 4: Results of the WSI inference strategy. In (a) the source image. In (b) the result using a naive tile-by-tile strategy. A clear artifact is present between the top and bottom tile. In (c) the tiling artifact is eliminated by using the overlapping tiling strategy.

for testing the effectiveness of stain transformation on segmentation performance. Per slide, 1-2 regions of interest were picked in which the selected renal structures were exhaustively annotated. Annotations for both datasets were made by a technician with experience in renal histopathology and checked by an experienced nephropathologist.

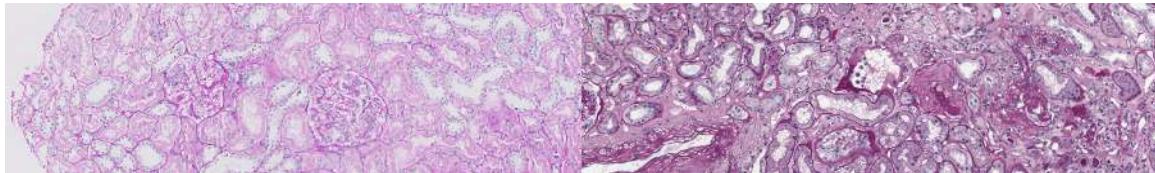


Figure 5: A tissue example from both data-centers, illustrating the color differences of the stains. Left: AMC. Right: RUMC

**Performance measures:** We use the structural similarity (SSIM) index to assess the benefit of our changes to the default cycleGAN architecture. SSIM is a perception-based metric that quantifies image degradation as change in structural information (Wang et al., 2004). SSIM is most commonly used in stain conversion approaches where there is a lack of paired tissue, as it compares the structure of images while largely disregarding the color scheme. We used  $C_1 = 0.01$ ,  $C_2 = 0.03$  and a window size of seven in our calculations. We use the network with the highest SSIM to perform the stain transformation.

Due to the lack of paired data, we can't use simple statistics like mean-squared difference to assess the quality of transformation. Instead, we compare the color histograms of synthetic and original stained patches. For this we use the Wasserstein distance between the histograms averaged across the RGB channels (Ling and Okada, 2007).

Last, to assess the segmentation network performance, we calculate the Dice coefficients on the ten annotated slides from the AMC dataset and the forty slides from the RUMC dataset, calculating the weighted average across the different classes. We report the average score over the ten slides, the standard deviation between the slide scores and the highest and lowest scored slides.

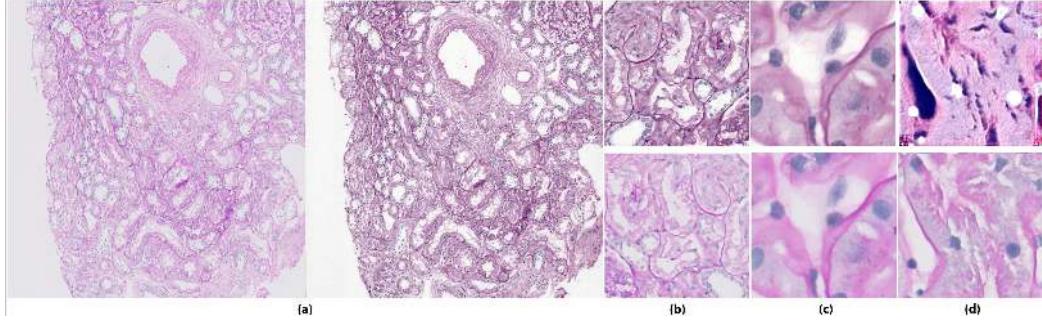


Figure 6: Samples of stain transformation by the network at increasing levels of zoom (a,b,c). At (d) an example of a failure run is given where the network learned to 'invert' the tissue. This problem was solved by adding the identity loss (1).

## 4. Results

**Performance of the stain transformation:** We present transformed samples in Figure 6. Additionally, randomly sampled tiles are shown in Appendix A. We report the SSIM values of our networks in Table 1, together with the histogram Wasserstein distances. As our network with 3 max-pooling layers obtained the best SSIM, we used it for the rest of the experiments.

**Comparison of segmentation performance with and without stain transfer:** The Dice coefficients obtained after segmentation are reported in Table 2. A table with scores per class is added in Appendix C. Qualitative results on the AMC dataset are shown in Figure 7, with additional examples in Appendix B.

## 5. Discussion

As expected, when applying the network trained without data augmentation on the non-transformed AMC data, the network fails with a Dice coefficient of 0.36. The segmentation network trained with data augmentation was able to achieve good results on the AMC dataset, improving the Dice coefficient from 0.36 to 0.78. However, there is still a performance gain when adding stain transformation on top of augmentation, increasing the average Dice coefficient from 0.78 to 0.85. This might indicate that not all the stain variation can be captured with only data augmentation and that the cycleGAN is better able to model non-linear stain variations. Interestingly, when stain transformation is applied, the average Dice coefficient is the same, regardless of whether the segmentation network was trained with or without augmentation. This provides evidence for when deploying these networks, re-training with data augmentation in case of protocol or scanner changes is not needed for algorithms downstream of the stain transformation cycleGAN.

The segmentation performance on the RUMC dataset shows a similar pattern. As we trained the segmentation networks with very few annotations, the overall scores expectantly turned out lower (Table 2, RUMC coefficients). Using either stain transformation or augmentation increased the Dice coefficient from 0.46 to 0.71. Using both techniques together gives a slight edge, increasing the score to 0.73. This supports our hypothesis that segmentation benefits from both augmentation and

stain transformation, combining non-linear and linear stain variations. Future research with more datasets will turn out whether the AMC dataset was an anomaly considering there was no increase in performance when using both augmentation and stain transformation. Over the two datasets we can conclude that stain transformation is at least as useful as augmentation.

There is no paired data available for our datasets, preventing the use of straightforward performance measures to quantitatively assess transformation quality (e.g. mean-squared difference). Instead, we opted to use the Wasserstein distance between the color histograms of the stains, to show that the color distributions of our transformed AMC slides are similar to the original RUMC slides.

We used the SSIM to show that the structural integrity of the original slides was not tampered with by the cycleGAN, demonstrating that our modifications score slightly better than the original cycleGAN architecture, while using less parameters. We think that the SSIM and Wasserstein distance on color histograms nicely complement each other, where the first quantifies the structure integrity and the second compares the color distributions.

In future work it would be valuable to assess our method on paired data to better quantitatively assess the performance of the stain transformation. This can, for example, be done by performing staining/re-staining. In this approach, a slide is cleared after the initial staining and scanning and then re-stained and scanned at a different site. Furthermore, we would like to investigate whether our approach directly translates to other types of stains, for example H&E or immunohistochemical stains. Finally, comparing different stain transformation methods, both other cycleGAN and classical machine learning approaches, will be an interesting venue to explore in future research.

## References

- Babak Ehteshami Bejnordi, Geert Litjens, Nadya Timofeeva, Irene Otte-Höller, André Homeyer, Nico Karssemeijer, and Jeroen AWM van der Laak. Stain specific standardization of whole-slide histopathological images. *IEEE transactions on medical imaging*, 35(2):404–415, 2016.
- Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- Francesco Ciompi, Oscar Geessink, Babak Ehteshami Bejnordi, Gabriel Silva de Souza, Alexi Baidoshvili, Geert Litjens, Bram van Ginneken, Iris Nagtegaal, and Jeroen van der Laak. The importance of stain normalization in colorectal tissue classification with convolutional networks. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 160–163. IEEE, 2017.
- Joseph Paul Cohen, Margaux Luck, and Sina Honari. How to cure cancer (in images) with unpaired image translation. In *International Conference on Medical Imaging with Deep Learning (MIDL 2018) – Abstract track*, 2018.
- Michael Gadermayr, Vitus Appel, Barbara M Klinkhammer, Peter Boor, and Dorit Merhof. Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 165–173. Springer, 2018.

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.
- Adnan Mujahid Khan, Nasir Rajpoot, Darren Treanor, and Derek Magee. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6):1729–1738, 2014.
- Haibin Ling and Kazunori Okada. An efficient earth mover’s distance algorithm for robust histogram comparison. *IEEE transactions on pattern analysis and machine intelligence*, 29(5):840–853, 2007.
- Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. doi: 10.23915/distill.00003. URL <http://distill.pub/2016/deconv-checkerboard>.
- Yair Rivenson, Hongda Wang, Zhensong Wei, Yibo Zhang, Harun Gunaydin, and Aydogan Ozcan. Deep learning-based virtual histology staining using auto-fluorescence of label-free tissue. *arXiv preprint arXiv:1803.11293*, 2018.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. *arXiv preprint arXiv:1804.01601*, 2018.
- David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.
- Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016.
- Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.

Table 1: The leftmost table shows the structural similarity between the original AMC slides and the transformed AMC slides. We also report the amount of parameters that the networks used. The rightmost table reports the average Wasserstein distances (WD) of the averaged color histograms. We compare: transformed AMC slides vs. RUMC slides, original AMC slides vs. RUMC, RUMC vs. RUMC.

	Param. (M)	SSIM		WD
cycleGAN-baseline	2.8	0.83		
Our approach (depth 1)	0.11	0.73		
Our approach (depth 2)	0.45	0.75		
Our approach (depth 3)	2.0	0.85		
Conv. AMC vs RUMC			3363	
Orig. AMC vs RUMC			14923	
RUMC vs. RUMC			4594	

Table 2: Dice coefficient of the segmentation of both datasets. Additionally, we show the highest and lowest scored slides and the standard deviation.

Experiment		Dice coefficient AMC				Dice coefficient RUMC			
Augmentations	Stain transformed	Mean	Std	Min	Max	Mean	Std	Min	Max
x	x	0.36	0.21	0.09	0.65	0.46	0.12	0.15	0.78
x	✓	<b>0.85</b>	0.06	0.69	0.91	0.71	0.12	0.34	0.87
✓	x	0.78	0.08	0.65	0.87	0.71	0.10	0.44	0.86
✓	✓	<b>0.85</b>	0.05	0.72	0.91	<b>0.73</b>	0.11	0.37	0.87

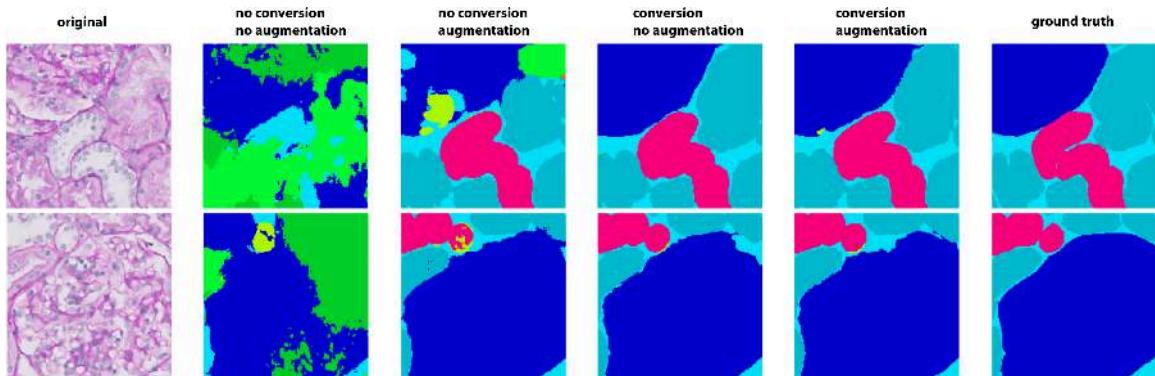


Figure 7: Samples from our segmentation results with and without augmentations and stain transformation.

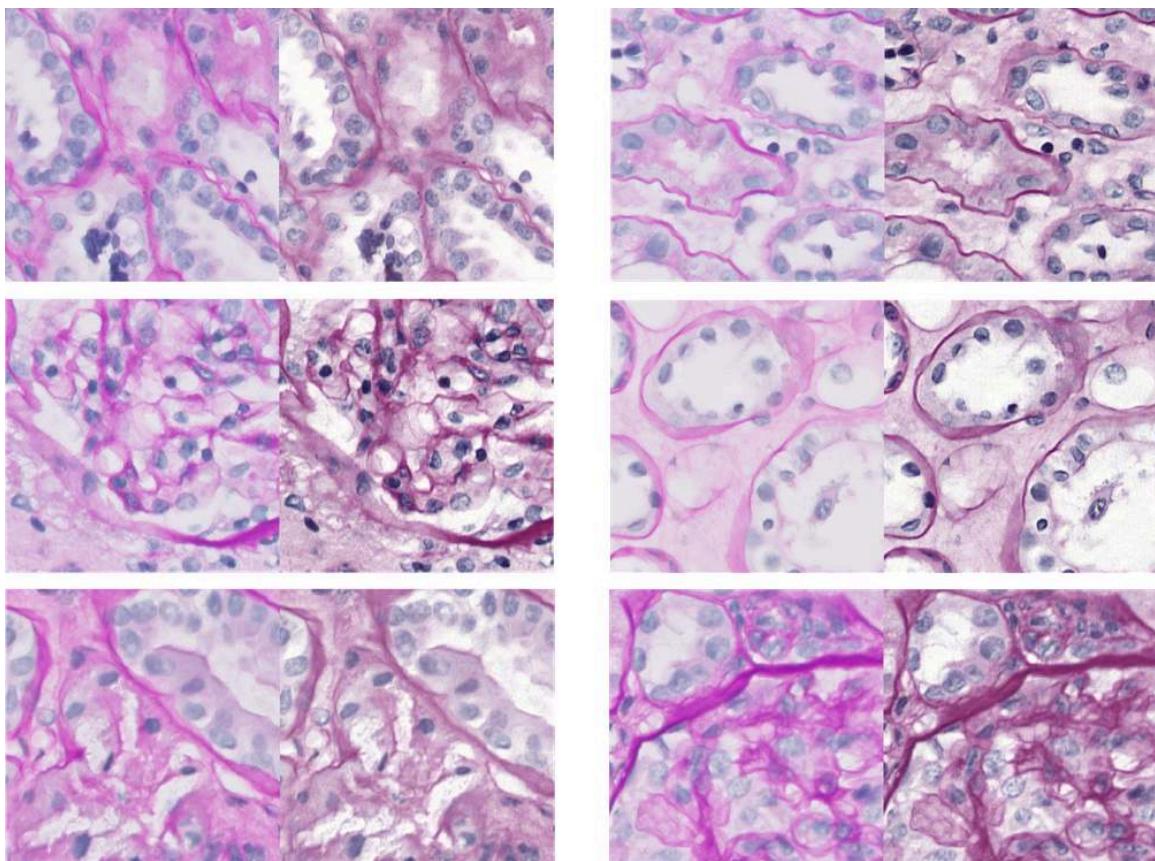
**Appendix A. Randomly sampled patches of real and artificially stained tissue**

Figure 8: Additional samples of stain transformation. The leftmost image of each tissue pair is from the original AMC stain, the rightmost the synthetic RUMC stain.

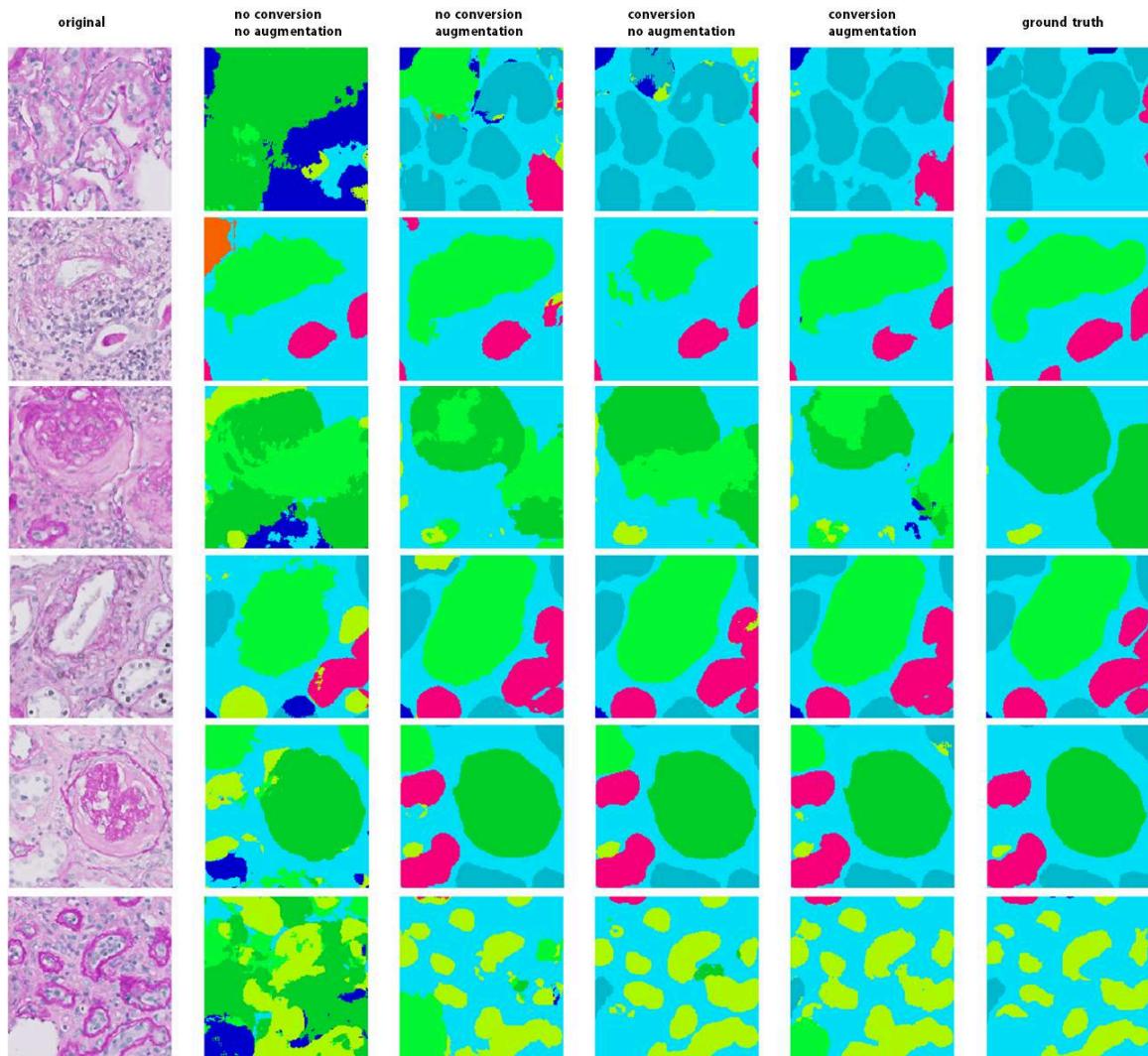
**Appendix B. Randomly sampled patches from the segmentation results**

Figure 9: Additional samples of segmentation results.

### Appendix C. Performance of the segmentation on the AMC dataset by class

Table 3: Dice coefficients of Sclerotic Glomeruli, Empty Glomeruli and Atrophic tubuli turned out considerably lower due to their low annotation count.

	aug, no conv	no aug, no conv	no aug, conv	aug, conv
Arteries	0.32	0.26	0.50	<b>0.51</b>
Atrophic tubuli	0.16	0.12	0.18	<b>0.19</b>
Background	0.79	0.49	<b>0.82</b>	<b>0.82</b>
Distal tubuli	0.64	0.23	<b>0.73</b>	0.71
Empty glomeruli	0.17	0.06	<b>0.24</b>	0.19
Glomeruli	0.78	0.45	0.88	<b>0.92</b>
Proximal tubuli	0.77	0.27	<b>0.85</b>	<b>0.85</b>
Sclerotic glomeruli	0.13	0.12	<b>0.32</b>	0.30

# Learning joint lesion and tissue segmentation from task-specific hetero-modal datasets

**Reuben Dorent<sup>1</sup>**

REUBEN.DORENT@KCL.AC.UK

**Wenqi Li<sup>1</sup>**

WENQI.LI@KCL.AC.UK

**Jinendra Ekanayake<sup>2</sup>**

J.EKANAYAKE@UCL.AC.UK

**Sebastien Ourselin<sup>1</sup>**

SEBASTIEN.OURSELIN@KCL.AC.UK

**Tom Vercauteren<sup>1</sup>**

TOM.VERCAUTEREN@KCL.AC.UK

<sup>1</sup> School of Biomedical Engineering & Imaging Sciences, King’s College London, London, UK

<sup>2</sup> Wellcome / EPSRC Centre for Interventional and Surgical Sciences, UCL, London, UK

## Abstract

Brain tissue segmentation from multimodal MRI is a key building block of many neuroscience analysis pipelines. It could also play an important role in many clinical imaging scenarios. Established tissue segmentation approaches have however not been developed to cope with large anatomical changes resulting from pathology. The effect of the presence of brain lesions, for example, on their performance is thus currently uncontrolled and practically unpredictable. Contrastingly, with the advent of deep neural networks (DNNs), segmentation of brain lesions has matured significantly and is achieving performance levels making it of interest for clinical use. However, few existing approaches allow for jointly segmenting normal tissue and brain lesions. Developing a DNN for such joint task is currently hampered by the fact that annotated datasets typically address only one specific task and rely on a task-specific hetero-modal imaging protocol. In this work, we propose a novel approach to build a joint tissue and lesion segmentation model from task-specific hetero-modal and partially annotated datasets. Starting from a variational formulation of the joint problem, we show how the expected risk can be decomposed and optimised empirically. We exploit an upper-bound of the risk to deal with missing imaging modalities. For each task, our approach reaches comparable performance than task-specific and fully-supervised models.

**Keywords:** joint learning, lesion segmentation, tissue segmentation, hetero-modality, weakly-supervision

## 1. Introduction

Traditional approaches for tissue segmentation used in brain segmentation software packages such as FSL (Jenkinson et al., 2012), SPM (Ashburner and Friston, 2000) or NiftySeg (Cardoso et al., 2015) are based on subject-specific optimisation. FSL and SPM fit a Gaussian Mixture Model to the MR intensities using either a Markov Random Field (MRF) or tissue prior probability maps as regularisation. Alternatively, multi-atlas methods rely on label propagation and fusion from multiple fully-annotated images, i.e. atlases, to the target image (Iglesias and Sabuncu, 2015). These methods typically require extensive pre-processing, e.g. skull stripping, correction of bias field or registration. They are also often time-consuming, and are inherently only adapted for brains devoid of large anatomical changes induced by pathology. Indeed, it has been showed that the presence of lesions distorts the registration output (Sdika and Pelletier, 2009). Similarly, lesions introduce a

bias in the MRF. This leads to a performance degradation in presence of lesions for brain volumes measurement ([Battaglini et al., 2012](#)) and any subsequent analysis.

While quantitative analysis is expected to play a key role in improving the diagnosis and follow-up evaluations of patients with brain lesions, current tools mostly focus on the lesions themselves. Existing quantitative neuroimaging approaches allow the extraction of imaging biomarkers such as the largest diameter, volume, and count of the lesions. Thus, automatic segmentation of the lesions promises to speed up and improve the clinical decision-making process but more refined analysis would be feasible from tissue classification and region parcellation. As such, although very few works have addressed this problem yet, a joint model for lesion and tissue segmentation is expected to bring significant clinical impact.

Deep Neural Networks (DNNs) became the state-of-the-art for most of the segmentation tasks and one would now expect to train a joint lesion and tissue segmentation algorithm. Yet, DNNs require a large amount of annotated data to be successful. Existing annotated databases are usually task-specific, i.e. providing either scans with brain tissue annotations for patients/controls devoid of large pathology-induced anatomical changes, or lesion scans with only lesion annotations. For this reason, the imaging protocol used for the acquisition also typically differs from one dataset to another. Indeed, tissue segmentation is usually performed on T1 scans, unlike lesion segmentation which normally also encompasses Flair ([Barkhof and Scheltens, 2002](#)). Similarly, the resolution and contrast among databases may also vary. Given the large amount of resources, time and expertise required to annotate medical images, given the varying imaging requirement to support each individual task and given the availability of task-specific databases, it is unlikely that large databases for every joint problem, such as lesion and tissue segmentation, will become available for research purposes. There is thus a need to exploit task-specific databases to address joint problems. Learning a joint model from task-specific hetero-modal datasets is nonetheless challenging. This problems lies at the intersection of Multi-Task Learning, Domain Adaptation and Weakly Supervised Learning with idiosyncrasies making individual methods from these underpinning fields insufficient to address it completely.

Multi-Task Learning (MTL) aims to perform several tasks simultaneously by extracting some form of common knowledge or representation and introducing a task-specific back-end. When relying on DNN for MTL, usually the first layers of the network are shared, while the top layers are task-specific. The global loss function is often a sum of task-specific loss functions with manually tuned weights. Recently, [Kendall and Gal \(2017\)](#) proposed a Bayesian parameter-free method to estimate the MTL loss weights and [Bragman et al. \(2018\)](#) extended it to spatially adaptive task weighting and applied it to medical imaging. In addition to arguably subtle differences between MTL and joint learning discussed in more depth later, MTL approaches do not provide any mechanism for dealing with hetero-modal datasets and changes in imaging characteristics across task-specific databases.

Domain Adaptation (DA) is a solution for dealing with heterogeneous datasets. The main idea is to create a common feature space for the two sets of scans. [Csurka \(2017\)](#) proposed an extensive comparison of these methods in deep learning. Learning from hetero-modal datasets could be consider as a particular case of DA. [Havaei et al. \(2016\)](#) proposed a network architecture for dealing with missing modalities. However, DA methods focus on solving a single task and rely on either fully-supervised approaches or unsupervised adaptation as done by [Kamnitsas et al. \(2017\)](#).

Weakly-supervised Learning (WSL) deals with missing, inaccurate, or inexact annotations. Our problem is a particular case of learning with missing labels since each dataset provide a set of labels and the two sets are disjoint. [Li and Hoiem \(2017\)](#) proposed a method to learn a new task from a

model trained on another task. This method combines DA through transfer learning and MTL. At the end, two models are created: one for the first task and one for the second one. Kim et al. (2018) extent this approach by using a knowledge distillation loss in order to create a unique joint model. This aims to alternatively learn one task without forgetting the other one. The WSL problem was thus decomposed into a MTL problem with similar limitations for our specific use case.

The contributions of this work are four-fold. First we propose a joint model that performs tissue and lesion segmentation as a unique joint task and thus exploits the interdependence between lesion and tissue segmentation tasks. Starting from a variational formulation of the joint problem, we exploit the disjointness of the label sets to propose a practical decomposition of the joint loss. Secondly, we introduce feature channel averaging across modalities to adapt existing networks for our hetero-modal problem. Thirdly, we develop a new method to minimise the expected risk under the constraint of missing modalities. Relying on reasonable assumptions, we show that the expected risk can be further decomposed and minimised via a tractable upper bound. To our knowledge, no such optimisation method for missing modalities in deep learning has been published before. Finally, we evaluate our framework for white matter lesions and tissue segmentation. We demonstrate that our joint approach can achieve, for each individual task, similar performance compared to a task-specific baseline. Albeit relying on different annotation protocols, results using a small fully-annotated joint dataset demonstrate efficient generalisability.

## 2. Tissue and lesion segmentation as a single task

In order to develop a joint model, we propose a mathematical variational formulation of the problem and a method to optimise it empirically.

### 2.1. Formal problem statement

Let  $\mathbf{x} = (x^1, \dots, x^M) \in \mathcal{X} = \mathbb{R}^{N \times M}$  be a vectorized multimodal image and  $y \in \mathcal{Y} = \{0, \dots, C\}^N$  its associated segmentation map.  $N$ ,  $M$  and  $C$  are respectively the number of voxels, modalities and classes. Our goal is to determine a predictive function  $h_\theta : \mathcal{X} \mapsto \mathcal{Y}$  that minimises the discrepancy between the ground truth label vector  $y$  and the prediction  $h_\theta(\mathbf{x})$ . Let  $\mathcal{L}$  be a loss function that computes this discrepancy. Following the formalism used by Bottou et al. (2018), given a probability distribution  $\mathcal{D}$  over  $(\mathcal{X}, \mathcal{Y})$  and random variables  $(X, Y)$  under this distribution, we want to find  $\theta^*$  such that:

$$\theta^* = \operatorname{argmin}_\theta \mathbb{E}_{(X,Y) \sim \mathcal{D}} [\mathcal{L}(h_\theta(X), Y)] \quad (1)$$

Let  $\mathcal{C}_t$ ,  $\mathcal{C}_l$  and  $0$  be respectively the tissue classes, the lesion classes and the background class. Since  $\mathcal{C}_t$  and  $\mathcal{C}_l$  are disjoint, the segmentation map  $y$  can be decomposed into two segmentation maps  $y_i = y_i^t + y_i^l$  with  $y_i^t \in \mathcal{C}_t \cup \{0\}$ ,  $y_i^l \in \mathcal{C}_l \cup \{0\}$ , as shown in Figure 1.

Let's assume that the loss function  $\mathcal{L}$  can also be decomposed into a tissue loss function  $\mathcal{L}^t$  and a lesion loss function  $\mathcal{L}^l$ . This is common for multi-class segmentation loss functions in particular for those with *one-versus-all* strategies (e.g. Dice loss, Jaccard loss):

$$\mathcal{L}(h_\theta(X), Y) = \mathcal{L}^t(h_\theta(X), Y^t) + \mathcal{L}^l(h_\theta(X), Y^l) \quad (\mathbf{H}_1)$$

Then, Equation (1) can be rewritten as:

$$\theta^* = \operatorname{argmin}_\theta \underbrace{\mathbb{E}_{(X,Y) \sim \mathcal{D}} [\mathcal{L}^t(h_\theta(X), Y^t)]}_{\mathcal{R}^t(\theta)} + \underbrace{\mathbb{E}_{(X,Y) \sim \mathcal{D}} [\mathcal{L}^l(h_\theta(X), Y^l)]}_{\mathcal{R}^l(\theta)} \quad (2)$$

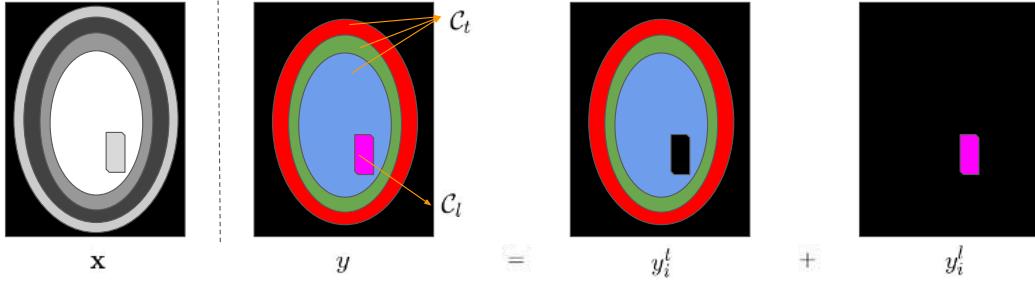


Figure 1: Decomposition of the label map into the sum of two segmentation maps.

## 2.2. On the distribution $\mathcal{D}$ in the context of heterogeneous databases

As we expect different distributions across heterogeneous databases, two probability distributions of  $(X, Y)$  over  $(\mathcal{X}, \mathcal{Y})$  can be distinguished: 1/ under  $\mathcal{D}_{control}$ ,  $(X, Y)$  corresponds to a multimodal scan and segmentation map of a patient without lesions. Note that although we use the term *control* for convenience, we expect to observe pathology with “diffuse” anatomical impact, e.g. from dementia. 2/ under  $\mathcal{D}_{lesion}$ ,  $(X, Y)$  corresponds to a multimodal scan and segmentation map of a patient with lesions.

Since traditional methods are not adapted in the presence of lesions, the most important and challenging distribution  $\mathcal{D}$  to address is the one for patients with lesions,  $\mathcal{D}_{lesion}$ . In the remainder of this work we thus assume that:

$$\mathcal{D} \triangleq \mathcal{D}_{lesion}. \quad (\mathbf{H}_2)$$

## 2.3. Hetero-modal network architecture

In order to learn from hetero-modal datasets, we need a network architecture that allows for missing modalities. We proposed an architecture inspired by HeMIS (Havaei et al., 2016) and HighResNet (Li et al., 2017) shown in Figure 2. Features of each modality are first extracted separately and are then averaged. The spatial resolution of the input and the output are the same. Dilated convolutions and residual connections are used to capture information at multiple scales and avoid the problem of vanishing gradients. This network with weights  $\theta$  is used to capture the predictive function  $h_\theta$ .

## 2.4. Upper-bound for the tissue expected risk $\mathcal{R}^t$

Although thanks to its hetero-modal architecture,  $h_\theta$  may now handle inputs with varying number of modalities, the current decomposition of (1) assumes that all the modalities of  $X$  are available for evaluating the loss. In our scenario, we have only access to T1 control scans with tissue annotations or T1 and Flair scans with only lesion annotations. Consequently, as we do not have any T1 and Flair images with tissue annotations, and as evaluating a loss with missing modalities would lead to a bias, estimating  $\mathcal{R}^t$  is not straightforward.

In this section we propose an upper-bound of  $\mathcal{R}^t$  using T1 control images with tissue annotations and outputs from the network. Let’s assume that the loss function  $\mathcal{L}^t$  satisfies the triangle inequality (e.g. Jaccard loss):

$$\forall (a, b, c) \in \mathcal{Y}^3 : \mathcal{L}^t(a, c) \leq \mathcal{L}^t(a, b) + \mathcal{L}^t(b, c) \quad (\mathbf{H}_3)$$

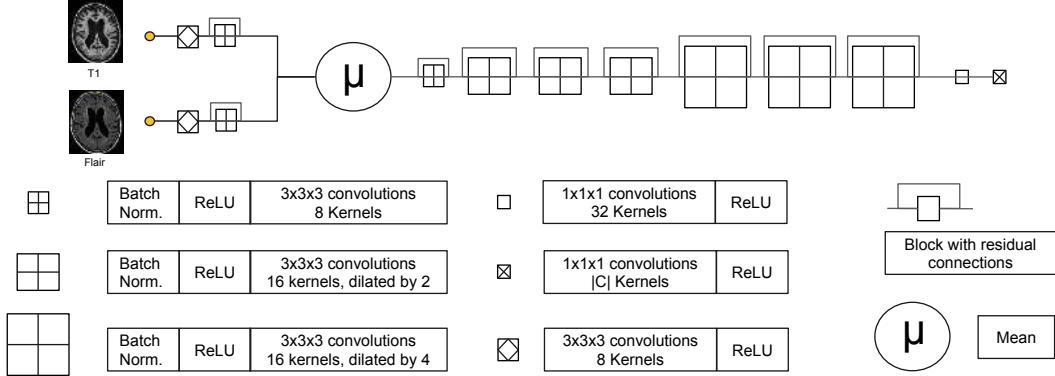


Figure 2: The proposed network architecture: a mix between HighResNet and HeMIS. To avoid cluttering, only one of the three convolution blocks is shown in the residual blocks.

Let  $p$  denote the projection of  $\mathbf{x}$  (will all the modalities) to the T1 modality,  $p : \mathbf{x} = (x^{\text{T1}}, x^{\text{Flair}}) \mapsto x^{\text{T1}}$ . Under  $(\mathbf{H}_3)$ ,  $\mathcal{L}^t$  satisfies the following inequality:

$$\mathcal{L}^t(h_\theta(X), Y^t) \leq \mathcal{L}^t(h_\theta(X), h_\theta(p(X))) + \mathcal{L}^t(h_\theta(p(X)), Y^t) \quad (3)$$

In combination with  $(\mathbf{H}_2)$ , this leads to:

$$\mathcal{R}^t(\theta) \leq \mathbb{E}_{(X,Y) \sim \mathcal{D}_{\text{lesion}}} [\mathcal{L}^t(h_\theta(X), h_\theta(p(X)))] + \mathbb{E}_{(X,Y) \sim \mathcal{D}_{\text{lesion}}} [\mathcal{L}^t(h_\theta(p(X)), Y^t)] \quad (4)$$

The decomposition in (4) requires comparison of inference done from T1 inputs, i.e.  $h_\theta(p(X))$  with ground truth tissue maps  $Y^t$ . While this provides a step towards a practical evaluation of  $\mathcal{R}^t$ , we still face the challenge of not having tissue annotations  $Y^t$  under  $\mathcal{D}_{\text{lesion}}$ . Let us further assume that the restriction of the distributions  $\mathcal{D}_{\text{lesion}}$  and  $\mathcal{D}_{\text{control}}$  to the parts of the brain not affected by lesions are the same, i.e.:

$$\forall i \in \{1 \dots N\} P_{\mathcal{D}_{\text{lesion}}}(x_i, y_i | y_i \in \mathcal{C}_{\text{tissue}}) = P_{\mathcal{D}_{\text{control}}}(x_i, y_i | y_i \in \mathcal{C}_{\text{tissue}}) \quad (\mathbf{H}_4)$$

By combining  $(\mathbf{H}_3)$  and  $(\mathbf{H}_4)$ , an upper bound of  $\mathcal{R}^t$  can be provided as:

$$\mathcal{R}^t(\theta) \leq \underbrace{\mathbb{E}_{(X,Y) \sim \mathcal{D}_{\text{lesion}}} [\mathcal{L}^t(h_\theta(X), h_\theta(p(X)))]}_{\mathcal{R}_1^t(\theta)} + \underbrace{\mathbb{E}_{(X,Y) \sim \mathcal{D}_{\text{control}}} [\mathcal{L}^t(h_\theta(p(X)), Y^t)]}_{\mathcal{R}_2^t(\theta)} \quad (5)$$

As observed in (5), the upper-bound is the sum of the expected loss between the T1 scan outputs and the labels and the expected loss between the outputs using either one or two modalities as input. We emphasise that, to the best of our knowledge, this second loss term does not appear in existing heteromodal approaches such as HeMIS (Havaei et al., 2016).

## 2.5. Empirical estimation of the decomposed loss

As is the norm in data-driven learning, we do not have access to the true joint probabilities  $\mathcal{D}_{control}$  or  $\mathcal{D}_{lesion}$ . The common method is to estimate the expected risk using training samples. In our case, we have two hetero-modal training samples  $\mathcal{S}_{control}$  and  $\mathcal{S}_{lesion}$  with respectively tissue and lesion annotations. We can estimate the expected risks  $\mathcal{R}^l(\theta)$ ,  $\mathcal{R}_1^t(\theta)$ ,  $\mathcal{R}_2^t(\theta)$  by respectively using lesion segmentation outputs of lesion T1+Flair scans, tissue segmentation outputs from T1 and T1+Flair scans and tissue segmentation outputs of control T1 scans. Figure 3 illustrates the complete training procedure.

### 3. Experiments

While focusing on the white matter lesion and tissue segmentation problem, our goal in the following experiments is to predict six tissue classes (white matter, gray matter, basal ganglia, ventricles, cerebellum, brainstem), the white matter lesions and the background.

#### 3.1. Data

To demonstrate the feasibility of our joint learning approach, we used three sets of data.

**Lesion data  $S_{lesion}$ :** The White Matter Hyperintensities (WMH) database consists of 60 sets of brain MR images (T1 and Flair,  $M = 2$ ) with manual annotations of WMH (<http://wmh.isi.uu.nl/>). The data comes from three different institutes.

**Tissue data  $S_{control}$ :** Neuromorphometrics provided 32 T1 scans ( $M' = 1$ ) for MICCAI 2012 with manual annotations of 155 structures of the brain from which we deduct the six tissue classes. In order to have balance training datasets for the two types of segmentation, and similar to [Li et al. \(2017\)](#), we added 28 T1 control scans from the ADNI2 dataset with bronze standard parcellation of the brain structures computed with the accurate but time-consuming algorithm of [Cardoso et al. \(2015\)](#).

**Fully annotated data:** MRBrainS18 (<http://mrbrains18.isi.uu.nl/>) is composed of 30 sets of brain MR images with tissue and lesions manual annotations. Only 7 MR images are publicly available. We used this data only for evaluation and not for training. To be consistent with the lesion data, the cerebrospinal fluid is considered as background.

To satisfy the assumption **(H<sub>4</sub>)**, we resampled the data to  $1 \times 1 \times 3 \text{ mm}^3$ , used a histogram-based scale ([Milletari et al., 2016](#)) and a zero-mean unit-variance normalization.

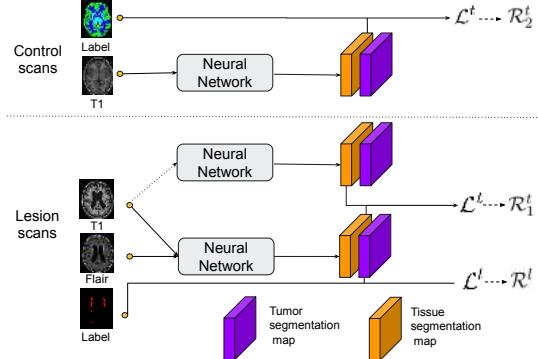


Figure 3: Procedure for estimating the expected risks  $\mathcal{R}^l$ ,  $\mathcal{R}_1^t$  and  $\mathcal{R}_2^t$ .

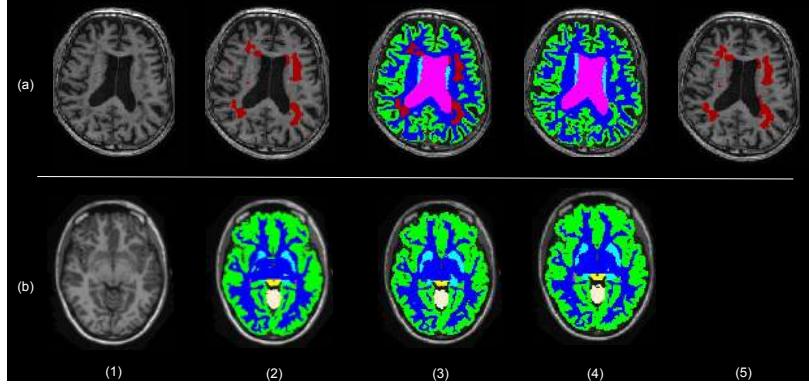


Figure 4: Segmentation results using our method and task-specific models. (1) axial slice from test image volumes from (a) WMH and (b) Neuromorphometrics, (2) manual annotations, (3) outputs from the joint learning model, (4) outputs from the tissue segmentation ( $N$ ) model, (5) outputs from the lesion segmentation ( $W$ ) model

### 3.2. Choice of the loss

We used the probabilistic version of the Jaccard loss for  $\mathcal{L}$ :

$$\mathcal{L}(h_\theta(\mathbf{x}), y) = 1 - \sum_{c \in C} \omega_c \frac{\sum_{j=1}^N g_{j,c} p_{j,c}}{\sum_{j=1}^N g_{j,c}^2 + p_{j,c}^2 - g_{j,c} p_{j,c}} \quad \text{such as } \sum_{c \in C} \omega_c = 1 \quad (6)$$

(**H<sub>1</sub>**) is satisfied because of the *one-versus-all* strategy, i.e. sum over the classes of a class-specific loss. In order to give the same weight to the lesion segmentation and the tissue segmentation, we choose for any tissue class  $c$ ,  $w_c = \frac{1}{16}$  and for the lesion class  $l$ ,  $w_l = \frac{1}{2}$ . While the triangle inequality holds for the Jaccard distance (Kosub, 2018), the proof that its probabilistic version also satisfies it, i.e. (**H<sub>3</sub>**), is left for future work.

### 3.3. Implementation details

We implemented our network in NiftyNet, a Tensorflow-based open-source platform for deep learning in medical imaging (Gibson et al., 2018). Convolutional layers are initialised such as He et al. (2015). The scaling and shifting parameters in the batch normalisation layers were initialised to 1 and 0 respectively. As suggested by (Ulyanov et al., 2016), we used instance normalization for inference. We used the Adam optimisation method (Kingma and Ba, 2014). The learning rate  $l_R$ ,  $\beta_1$ ,  $\beta_2$  were respectively set up to 0.005, 0.9 and 0.999. At each training iteration, we feed the network with one image from the tissue dataset and one from the lesion dataset.  $120 \times 120 \times 40$  sub-volumes were randomly sampled from the training data using an uniform sampling for the tissue data and a weighted sampling based on dilated lesions maps for the lesion data. The models were trained until we observed a plateau in performance on the validation set. We experimentally found that the inter-modality loss has to be skipped for the first (5000) iterations. We randomly spitted the data into 70% for training, 10% for validation and 20% for testing for each of the 4 folds.

### 3.4. Results for the joint learning model

**Joint learning versus single task learning** First, we compare individual models to the joint model using our approach. The lesion segmentation ( $W$ ) model was trained on WMH dataset with the lesion annotations, the tissue segmentation ( $N$ ) on Neuromorphometrics dataset with the tissue annotations, and our method ( $W+N$ ) on WMH and Neuromorphometric datasets with their respective set of annotations. The similarity between the prediction and the ground truth is computed using the Dice Similarity Coefficient (DSC) for each class. Table 1 and Figure 4 show the results of these models on test images. The joint model and single task models achieve comparable performance. This suggest that learning from hetero-modal datasets via our method does not degrade the task-specific performance. Moreover, we observe in Figure 4 that the tissue knowledge learnt from T1 scans has been well generalised to multi-modal scans.

Table 1: Comparison between the lesion segmentation model  $W$ , the tissue segmentation model  $N$ , the fully-supervised model ( $M$ ), a traditional approach ( $SPM$ ) and our joint model ( $W+N$ ). Dice Similarity Coefficients (%) has been computed.

	Neuromorphometrics			WMH			MRBrainS18		
	$N$	$M$	$W+N$	$W$	$M$	$W+N$	$SPM$	$M$	$W+N$
Gray matter	88.5	42.0	89.4				76.5	83.3	79.4
White matter	92.4	56.7	92.8				75.7	85.9	85.4
Brainstem	93.4	20.0	93.1				76.5	92.3	72.3
Basal ganglia	86.7	41.2	87.2				74.7	79.1	75.3
Ventricles	90.7	24.5	91.6				80.9	91.0	91.7
Cerebellum	92.5	43.7	94.9				89.4	91.8	90.8
White matter lesion				61.9	50.6	59.9	40.8	53.5	53.7

**Joint model versus fully-supervised model** In this section, we compare our method ( $W+N$ ) to the fully-supervised ( $M$ ) model trained on MRBrainS18 using both tissue and lesion annotations. We evaluated the performance on the three datasets. On the one hand, we submitted our models to the online challenge MRBrainS18. One of the major benefits of evaluating our method on a challenge is to directly benchmark our method with existing methods, in particular with traditional methods such as SPM (Ashburner and Friston, 2000). On the other hand, we compared the performance on the tissue and lesion datasets using either all the scans ( $M$ ) or the testing split ( $W+N$ ). The DSC was computed for each class and Table 1 show the results. First our model outperforms SPM on 6 of the 7 classes. Secondly, the two models achieve very similar performance on lesion segmentation. Concerning the tissue segmentation, as expected, each of the network outperforms on its training datasets. However, the fully supervised model doesn't generalise to Neuromorphometrics dataset. In contrast, the differences are smaller for the tissue segmentation classes on MRBrainS18. Especially, Figure 5 shows differences in the annotation protocol between MRBrainS18 and Neuromorphometrics for the white matter, brainstem and cerebellum and how it affects the predictions.

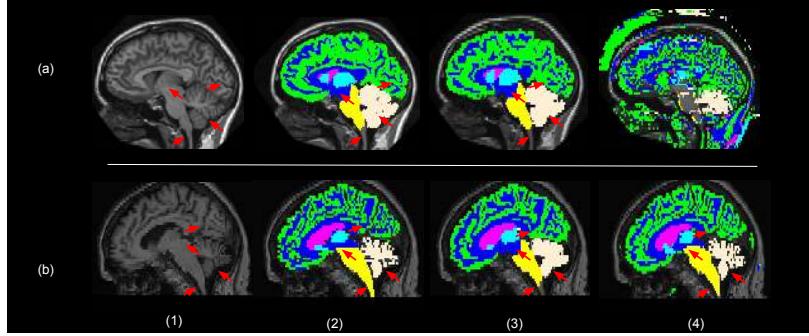


Figure 5: Annotation protocol comparison between scans from (a) Neuromorphometrics and (b) MRBrainS18. (1) sagital slice from test images volumes, (2) manual annotations, (3) outputs from our method ( $W+N$ ), (4) outputs from fully-supervised model ( $N$ ). Arrows show the protocol differences.

#### 4. Conclusion

We propose a joint model learned from hetero-modal datasets with disjoint heterogeneous annotations. Our approach is mathematically grounded, conceptually simple, new and relies on reasonable assumptions. We validated our approach by comparing our joint model with single-task learning models. We show that similar performance can be achieved for the tissue segmentation and lesion segmentation in comparison to task-specific baselines. Moreover, our model achieves comparable performance to a model trained on a small fully-annotated joint dataset. Our work shows that the knowledge learnt from one modality is preserved when more modalities are used as input. In the future, we will evaluate our approach on datasets with annotations protocols showing less variability. Furthermore, exploitation of recent techniques for domain adaptation could help us bridge the gap and improve the performance by helping to better satisfy some of our assumptions. Finally, we plan to integrate uncertainty measures in our framework as a future work. As one of the first work to methodologically address the problem of joint learning from hetero-modal datasets, we believe that our approach will help DNN make further impact in clinical scenarios.

#### Acknowledgments

This work was supported by the Wellcome Trust [203148/Z/16/Z] and the Engineering and Physical Sciences Research Council (EPSRC) [NS/A000049/1]. TV is supported by a Medtronic / Royal Academy of Engineering Research Chair [RCSR1819\7\34].

#### References

- John Ashburner and Karl J. Friston. Voxel-based morphometry – the methods. *Neuroimage*, 11(6): 805–821, 2000.
- Frederik Barkhof and Philip Scheltens. Imaging of white matter lesions. *Cerebrovascular Diseases*, 13(Suppl 2):21–30, 2002.

- Marco Battaglini, Mark Jenkinson, and Nicola De Stefano. Evaluating and reducing the impact of white matter lesions on brain volume measurements. *Human Brain Mapping*, 33(9):2062–2071, 2012.
- Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.
- Felix J. S. Bragman, Ryutaro Tanno, Zach Eaton-Rosen, Wenqi Li, David J. Hawkes, Sébastien Ourselin, Daniel C. Alexander, Jamie R. McClelland, and M. Jorge Cardoso. Uncertainty in multitask learning: Joint representations for probabilistic MR-only radiotherapy planning. In *Proceedings of the 21st International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'18)*, pages 3–11, 2018.
- M. Jorge Cardoso, Marc Modat, Robin Wolz, Andrew Melbourne, David Cash, Daniel Rueckert, and Sébastien Ourselin. Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion. *IEEE Transactions on Medical Imaging*, 34(9):1976–1988, September 2015.
- Gabriela Csurka. *A comprehensive survey on domain adaptation for visual applications*, pages 1–35. Springer International Publishing, 2017.
- Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I. Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, Tom Whyntie, Parashkev Nachev, Marc Modat, Dean C. Barratt, Sébastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. NiftyNet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, 158: 113–122, 2018.
- Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. HeMIS: Hetero-modal image segmentation. In *Proceedings of the 19th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'16)*, pages 469–477, 2016.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 25th International Conference on Computer Vision (ICCV'15)*, pages 1026–1034, December 2015.
- Juan Eugenio Iglesias and Mert R Sabuncu. Multi-atlas segmentation of biomedical images: a survey. *Medical Image Analysis*, 24(1):205–219, August 2015.
- Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. FSL. *Neuroimage*, 62(2):782–790, 2012.
- Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *Proceedings of the 20th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'17)*, pages 597–609, 2017.
- Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 5574–5584, 2017.

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, Youngjin Yoon, and In So Kweon. Disjoint multi-task learning between heterogeneous human-centric tasks. In *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (2018)*, pages 1699–1708, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).

Sven Kosub. A note on the triangle inequality for the Jaccard distance. *Pattern Recognition Letters*, December 2018.

Wenqi Li, Guotai Wang, Lucas Fidon, Sébastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. On the compactness, efficiency, and representation of 3D convolutional networks: Brain parcellation as a pretext task. In *Proceedings of Information Processing in Medical Imaging (IPMI'17)*, pages 348–360, 2017.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, November 2017.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

Michaël Sdika and Daniel Pelletier. Nonrigid registration of multiple sclerosis brain images using lesion inpainting for morphometry or lesion mapping. *Human Brain Mapping*, 30(4):1060–1067, 2009.

Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization, 2016. [arXiv:1607.08022](https://arxiv.org/abs/1607.08022).

# Unsupervisedly Training GANs for Segmenting Digital Pathology with Automatically Generated Annotations

**Michael Gadermayr<sup>1,2</sup>**

**Laxmi Gupta<sup>2</sup>**

<sup>1</sup> Salzburg University of Applied Sciences, Austria

<sup>2</sup> Institute of Imaging and Computer Vision, RWTH Aachen University

**Barbara M. Klinkhammer<sup>3</sup>**

**Peter Boor<sup>3</sup>**

<sup>3</sup> Institute of Pathology, University Hospital Aachen, RWTH Aachen University, Germany

**Dorit Merhof<sup>2</sup>**

## Abstract

Recently, generative adversarial networks exhibited excellent performances in semi-supervised image analysis scenarios. In this paper, we go even further by proposing a fully unsupervised approach for segmentation applications with prior knowledge of the objects' shapes. We propose and investigate different strategies to generate simulated label data and perform image-to-image translation between the image and the label domain using an adversarial model. For experimental evaluation, we consider the segmentation of the glomeruli, an application scenario from renal pathology. Experiments provide proof of concept and also confirm that the strategy for creating the simulated label data is of particular relevance considering the stability of GAN trainings.

## 1. Motivation

Due to the progressing dissemination of whole slide scanners generating large amounts of digital histological image data, image analysis in this field has recently gained significant importance (Hou et al., 2016; BenTaieb and Hamarneh, 2016; Gadermayr et al., 2018a; Valkonen et al., 2017; Veta et al., 2016; Gadermayr et al., 2017; Herve et al., 2011).

For segmentation applications, especially fully-convolutional networks proved to be highly effective tools (Ronneberger et al., 2015; BenTaieb and Hamarneh, 2016; Gadermayr et al., 2017). A major challenge in the field of digital pathology is given by a large set of different application scenarios as well as changing underlying data distributions which is due to inter-subject variability, different staining protocols and/or pathological modifications (Gadermayr et al., 2018b). Each individual application scenario therefore requires large amounts of annotated training data covering the prevalent variability. The acquisition of such large amounts of labeled training data, however, is typically time-consuming and cost-intensive and thereby constitutes a burden for the deployment of automated segmentation techniques.

For training state-of-the-art machine learning approaches such as fully-convolutional networks, data augmentation proved to be a highly powerful tool (Ronneberger et al., 2015; J. Ratner et al., 2017) to keep the amount of required training data decent. A limitation of data augmentation in combination with supervised learning approaches is given by the fact that often large non-annotated

data is available “for free” but is not utilized for training at all. Particularly in the fields of medicine, such as digital pathology, huge amounts of digital image data are routinely captured without any (additional) effort whereby a complete annotation of all data is definitely not feasible. In order to take advantage of non-annotated data as well, dedicated semi-supervised segmentation approaches relying on adversarial models were recently proposed (Kozí Nski et al., 2017; Isola et al., 2017; Hung et al., 2018).

Adversarial models were also developed for the field of image-to-image translation (Johnson et al., 2016; Zhu et al., 2017). Recently, the so-called cycleGAN (Zhu et al., 2017) was introduced which eliminates the restriction of corresponding image pairs for training. This architecture can also be utilized for means of unsupervised domain adaptation (Chartsias et al., 2017; Wolterink et al., 2017; Gadermayr et al., 2018a). The domain adaptation in these cases is performed on image-level, i.e. “fake” images showing similar characteristics as the target domain samples are generated. This strategy is highly flexible as it can be combined with arbitrary further segmentation or classification approaches.

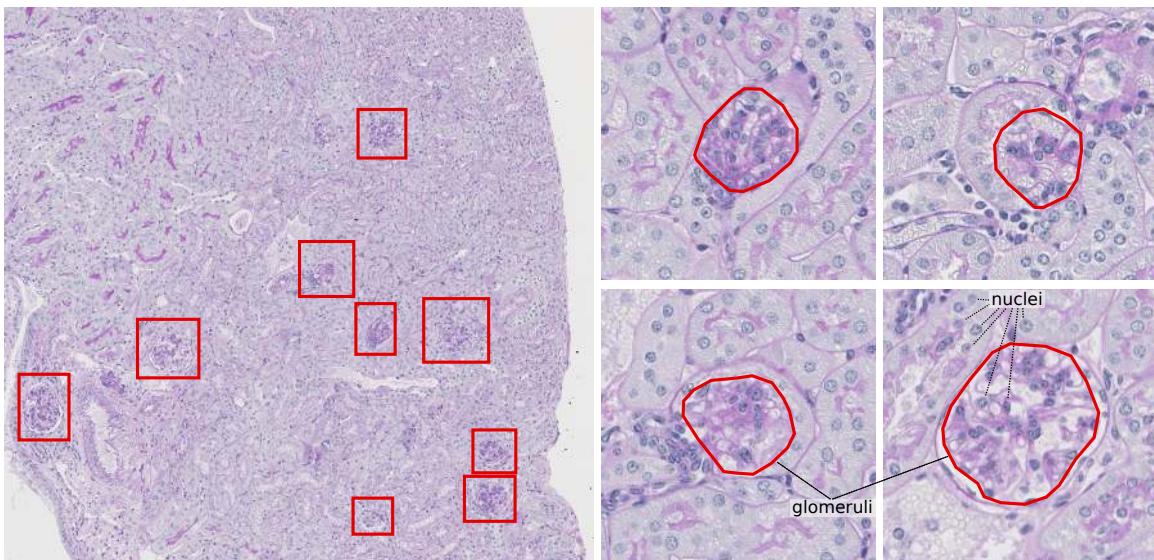


Figure 1: This illustration shows an extract of a renal whole slide image with marked glomeruli (left) as well as magnifications of single glomeruli showing precise manual annotations (right).

**Contribution:** We tackle the problem of acquiring labeled training data by proposing a framework completely bypassing the need for manually labeled objects. We focus on generating artificial annotations to perform image-to-image translation on unpaired data sets. In our experimental study, we investigate strategies for modeling the shape of the annotations and for modeling additional image information to facilitate training the translation networks. As application scenario, we consider a segmentation task from digital pathology, specifically the segmentation of the renal glomeruli (Gadermayr et al., 2017; Kato et al., 2015; Herve et al., 2011) (Fig. 1).

## 2. Methods

For the proposed method, we make use of an image-to-image translation approach. Specifically, we utilize a generative adversarial network (GAN) which facilitates training with unpaired data (Zhu et al., 2017). The four subnetworks consisting of two generators and two discriminators are optimized based on an adversarial loss as well as a cycle consistency criterion. This formulation does not require sample pairs, i.e. there is no need to obtain corresponding image samples for the two domains. Instead it is sufficient to collect a set of images, individually for each domain. If annotations are interpreted as label (e.g. binary) images, this approach can be utilized for segmentation applications as well. The architecture allows to perform training based on a set of images and a (non-corresponding) set of annotations as long as the annotations are realistic (i.e. the distribution matches the underlying distribution of real annotations).

The proposed method relies on an automated generation of realistic annotation images followed by training an image-to-image translation model which is finally able to convert original images to annotations. The procedure is based on the following assumptions: (1) we need to understand the underlying distribution of the annotation data and we need to be able to model this distribution (for details, see Sect. 2.1). (2) The unpaired image-to-image translation approach needs to be effectively applicable to translate between the image and the annotation domain. If a straight-forward translation is not effective, additional information can be added to the annotation domain to enhance the translation process (for details, see Sect. 2.2).

### 2.1. Annotation Model

In the considered application scenario (Fig. 1), the underlying distribution (assumption 1) of the objects-of-interest is rather basic and can thereby be approximately modeled quite well. The objects-of-interest show roundish shapes which are sparsely distributed over the kidney. For training we consider patches extracted from the whole slide images. We assume that the number of objects per patch can be approximated by a (quantized) Gaussian distribution  $G_{\#} \sim \mathcal{N}(\mu_g, \sigma_g^2)$ . The objects are uniformly distributed over the patch with one single further assumption that the objects may not overlap. For generating the annotations, we investigate two different approaches. Firstly, we consider the objects-of-interest as round objects (**Circular objects (C):**). The objects' radii  $r$  are randomly sampled from a Gaussian distribution  $R \sim \mathcal{N}(\mu_r, \sigma_r^2)$ . In a second configuration, we incorporate the fact that the objects-of-interest often show an elliptic shape (**Elliptic objects (E):**). To incorporate this knowledge,  $r_1$  is drawn from the same distribution as  $r$  and  $r_2 = r_1 + r_\delta$  where  $r_\delta$  models the eccentricity and is drawn from  $R_\delta \sim \mathcal{N}(0, \sigma_e^2)$ . A further rotation parameter  $\alpha$  is drawn from a uniform distribution in the interval  $[0, 2\pi]$ .

### 2.2. Image-to-Label Translation

The straightforward approach consists of adding either circles or ellipses as binary objects into two dimensional matrices which are interpreted as single channel images. However, for training the image-to-image translation approach, this setting can be highly challenging due to the loss criteria:

For training the GAN (Zhu et al., 2017), two generative models,  $F : \mathcal{X} \rightarrow \mathcal{Y}$  and  $G : \mathcal{Y} \rightarrow \mathcal{X}$  and two discriminators  $D_X$  and  $D_Y$  are trained optimizing the cycle consistency loss  $\mathcal{L}_c$

$$\begin{aligned} \mathcal{L}_c = & \mathbb{E}_{x \sim p_{data}(x)} [||G(F(x)) - x||_1] + \\ & \mathbb{E}_{y \sim p_{data}(y)} [||F(G(y)) - y||_1] \end{aligned} \tag{1}$$

and the adversarial loss  $\mathcal{L}_d$

$$\begin{aligned}\mathcal{L}_d = & \mathbb{E}_{x \sim p_{data}(x)} [\log(D_X(x)) + \log(1 - D_Y(F(x)))] + \\ & \mathbb{E}_{y \sim p_{data}(y)} [\log(1 - D_X(G(y))) + \log(D_Y(y))].\end{aligned}\quad (2)$$

$F$  and  $G$  try to generate fake images that look similar to real images, while  $D_X$  and  $D_Y$  aim to distinguish between translated and real samples. The generators aim to minimize this adversarial objective against the discriminators that try to maximize it.

Let  $X$  be the domain referring to the original images and let  $Y$  be the label domain. The cycle criterion requires that an annotation mask can be translated to an image by the generator  $G$ . The generator  $F$ , however, hides all low-level image details, such as nuclei and tubuli (Fig. 1) and only preserves the high-level shapes of the glomeruli. Based on these shapes only, it will not be able to reconstruct e.g. the nuclei at the right (i.e. the same) positions leading to a high cycle-consistency loss even though the images might look realistic. To take this into account, we propose and investigate a second setting simulating the nuclei exhibiting low-level information as well. As for the glomeruli, the number of nuclei is drawn from a (quantized) normal distribution  $N_\# \sim \mathcal{N}(\mu_n, \sigma_n^2)$ . They are uniformly distributed over the whole patch with the restriction that they may not overlap. Diameter is fixed to  $d_n$ . The additional binary matrix containing the nuclei is added as further image channel to the annotation image. This channel is only needed to train the GAN. For testing, this further channel is simply ignored. Whereas the setting incorporating only the target labels (i.e. the glomeruli) is referred to as single-class scenario, the setting also incorporating further low-level information is referred to as multi-class scenario. Finally, we identified four different settings: single-class circular objects (SC), single-class elliptical objects (SE), multi-class circular objects (MC) and multi-class elliptical objects (ME).

To facilitate learning, Gaussian random noise ( $\sigma_{f_n}$ ) is added to the annotation maps followed by the application of a Gaussian filter ( $\sigma_{f_s}$ ) to smooth the objects' borders in all settings.

### 2.3. Experimental Setting

Paraffin sections ( $1\mu m$ ) are stained with periodic acid-Schiff (PAS) reagent and counterstained with hematoxylin. Whole slides are digitalized with a Hamamatsu NanoZoomer 2.0HT digital slide scanner and a  $20\times$  objective lens. From each of the 23 WSIs overall, 200 patches with a size of  $500 \times 500$  pixels are randomly extracted (resulting in 4600 patches overall). For evaluation purpose, the WSIs are manually annotated under the supervision of a medical expert. Learning is performed in a transductive setting, i.e. both training and testing is executed on all patches. This does not introduce bias in this case, as no label data is used during training.

As large context is required to assess whether segmentations are realistic, a (rather low) resolution corresponding to a  $2.5\times$  magnification is utilized (original images downsampled by factor eight).

For image-to-image translation, we make use of the cycle GAN (Zhu et al., 2017). We rely on the provided pytorch reference implementation. Apart from the following changes, we use the proposed standard settings. As generator model, a residual network consisting of four blocks is utilized. As discriminator, we rely on the suggested patch-wise CNN with three layers (Zhu et al., 2017). Learning rate is fixed to  $10^{-6}$ , number of training epochs is set to 15 and batch size is set to one. The losses are equally weighted. For data augmentation, flipping, rotation and random cropping ( $384 \times 384$  pixel sub-patches) is performed.

The annotations are generated based on the following visually assessed parameters (we did not incorporate statistical information of the data set to avoid introducing significant supervision):  $\mu_g = 7$ ,  $\sigma_g = 2$ ,  $\mu_r = 18$ ,  $\sigma_r = 2$ ,  $\sigma_e = 2$ ,  $d_n = 4$ ,  $\mu_n = 5000$ ,  $\sigma_n = 50$ ,  $\sigma_{f_n} = 5$  and  $\sigma_{f_s} = 2$ .

For evaluation, we investigate two optimization strategies. The first strategy does not incorporate any optimization and we basically report the obtained segmentation performance after training for all 15 epochs. As GAN training is, in general, often unstable, we also optimize the epoch by separating the testing data set into one patch for optimization and the others for testing. We use only one patch for optimization because the approach is intended to be unsupervised.

Apart from pixel-level scores ( $F_1$ -score (F), precision (P), recall(R)), we also report the corresponding object-level scores ( $F_o$ ,  $P_o$ ,  $R_o$ ). That means, we distinguish between true positive objects (i.e. the distance between the center of a detected object and a real object is smaller than 10 pixels), objects which were missed and false positively detected objects.

All experiments are repeated four times. The obtained performances are compared with the a supervised fully-convolutional network ([Gadermayr et al., 2017](#)).

### 3. Results

Fig. 2 shows quantitative results for each of the four different settings. We investigate pixel-level as well as object-level scores. The left two columns show the testing pixel-level and object-level  $F_1$ -scores for different numbers of training epochs. The third column shows the scores obtained with cross validation (i.e. the epoch is optimized) and the last column shows the rates corresponding to training for 15 epochs without any further optimization.

Considering these results, we notice that the single-class settings (SC, SE) do not show any useful results. In case of elliptical shapes (SE), at least the best configuration exhibits acceptable outcomes, however, GAN training is highly unstable in this scenario. In case of the multi-class settings (MC, ME), we notice a more stable behavior, as in each repetition good scores are obtained after few training epochs. Mean pixel-level  $F_1$ -scores of 0.63 (MC) and 0.62 (ME) as well as mean object-level F-scores of 0.74 (MC and ME) are achieved. Convergence is obtained approximately after six epochs for both settings. We notice slightly higher precision than recall, especially on object-level. A further optimization of the number of the training epoch does not show a high influence.

The baseline results of the supervised approach are provided in Fig. 3. We obtain  $F_1$ -scores of 0.49, 0.65 and 0.71 on pixel level and 0.52, 0.68 and 0.76 on object-level for training with 2, 4 and 8 WSIs. We notice that the break-even point of the supervised approach is reached with approximately four fully-annotated training WSIs corresponding to roughly 500 annotated glomeruli. Considering the object-level scenario, the proposed method exhibits increased performances (comparable with the supervised method trained on eight WSIs).

We further investigated the annotation images with respect to the shape of the masks. Comparing the best fitting circle with the mask showed a mean  $F_1$ -score of 0.92 while for the best fitting ellipse, a  $F_1$ -score of 0.95 is obtained. The difference is statistically significant ( $p < 0.001$ ) and indicates that ellipses provide better approximations for the objects-of-interest.

Example output of the image-to-image translation process is provided in Fig. 4. With the single-class setting ((a)-(b)), we notice a tendency to segment vessel structures instead of the target objects. This is not the case if making use of the multi-class settings ((c)-(d)).

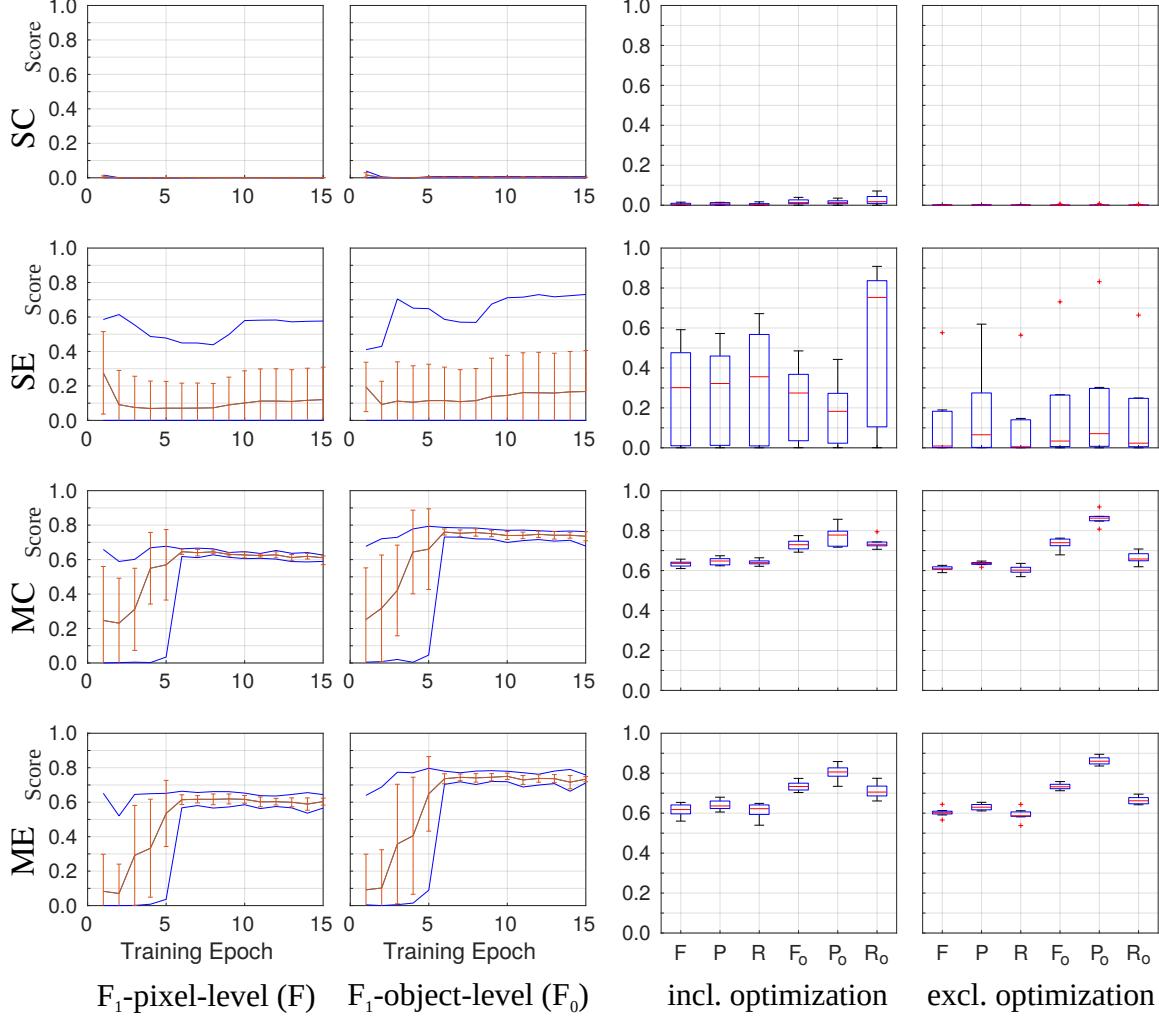


Figure 2: Experimental results for the four different settings (row 1 to row 4). The left columns show pixel- and object-level F<sub>1</sub>-scores for testing after varying number of training epochs. The right columns provide F<sub>1</sub>-scores (F), precision (P) and recall (R) as well as object-level measures (F<sub>0</sub>, P<sub>0</sub>, R<sub>0</sub>) for training for 15 epochs (excl. optimization) and for optimizing the epoch (incl. optimization).

#### 4. Discussion

In this work, we investigated a concept of fully-unsupervised learning for segmentation applications by making use of a GAN in combination with simulated annotation data.

We obtained highly divergent results for the four different settings. One substantial finding is that a simulation of the annotations of the objects-of-interest only (referred to as single-class scenario) is not sufficient to obtain proper segmentations of the glomeruli in the investigated unpaired image-to-image translation scenario. In the majority of attempts, an unwanted translation between the image and the label domain is observed. A major problem here is that a translation from the

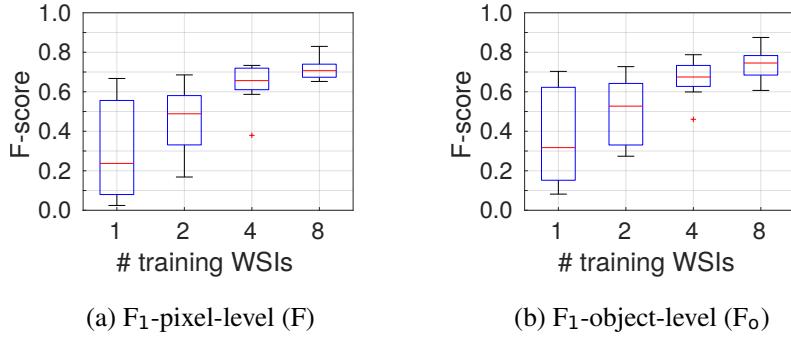


Figure 3: Baseline pixel-level (a) and object-level (b)  $F_1$ -scores indicating the segmentation performance of the supervised U-Net-based approach (Gadermayr et al., 2017) with variable numbers of fully-annotated training WSIs. One single training WSI contains on average 120 single objects.

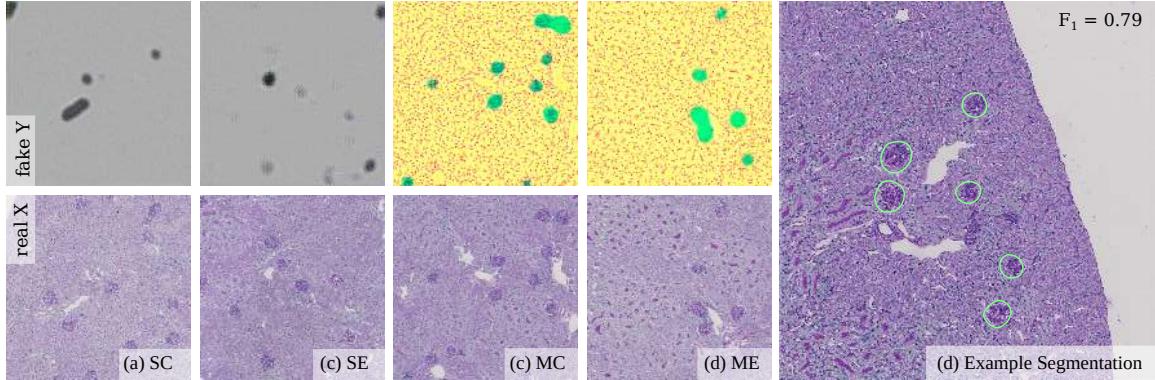


Figure 4: Qualitative results of the image translation process for the four different settings (a) – (d). While the setting SC and SE do not show any good segmentations, MC and ME perform similarly well. Subfigure (e) shows an example segmentation extracted from a fake image generated with setting MC.

label to the image domain cannot be performed which complicates the GAN training. The generator  $G$  in this case has no chance to place the low-level objects (here the nuclei) in a way that the cycle consistency loss can become small as the position of the nuclei cannot be effectively derived from the annotation image. This behavior can also be seen in the example reconstructed images where nuclei cannot be clearly detected (Fig. 4, third column). In the multi-class scenarios with added simulated nuclei during GAN training, these objects are maintained during the training cycles. That means, the nuclei are segmented during translation to the label domain followed by a reconstruction of the nuclei based on the label domain in case of the inverse mapping.

A further interesting finding is that the distribution of the shapes of the simulated objects does not have a major impact on final segmentation performance. We do not consider the single-class scenarios here as they showed either completely wrong or highly unstable performance. The multi-

class scenarios show similar performances for the setting based on circles and for the setting based on ellipses.

Considering the multi-class settings MC and ME, we assess the obtained segmentation performance as good and applicable for medical applications although the scores seem to be rather low. We need to mention here that this is on the one hand due to the fact that small objects are often not identified as glomeruli in the ground-truth but are detected by our approach. On the other hand, there are also small objects which are in the ground-truth but are not detected. Anyway, these objects are neglected by the medical experts and are thereby excluded from further analysis.

A comparison with a state-of-the-art supervised approach showed that the novel method is highly competitive. Especially the detection performance (indicated by the object-level  $F_1$ -scores) is outperformed by the supervised technique only if training is performed with a large amount of annotated data (specifically with eight WSIs corresponding to approx. 1000 single objects). Due to the stable training process, a “slightly-supervised” optimization of the training epoch is not required as the results are only marginally improved (Fig. 2, column 3 vs. column 4).

The most notable advantage, however, does not consist in high scores, but in a very high flexibility. The method can be easily adapted e.g. to other stains without a need for collecting novel annotated training data. An intrinsic limitation is certainly given regarding the shape of the objects-of-interest. While rather basic shapes can be easily modeled, complex or irregular shapes are either difficult or even impossible to model.

To conclude, we proposed and investigated a concept of fully-unsupervised learning for segmentation applications by making use of a GAN trained with real images and simulated annotations. The experimental results, in general highly promising, indicate that it is not crucial to accurately model the underlying shape as long as a good approximation is available. This is a highly relevant finding as the shapes of the objects-of-interest are often too complex to be modeled accurately. It is clearly more relevant to support the GAN to fulfill the cycle consistency criterion. Adding additional information to the label domain proved to be an effective way to facilitate the unpaired training process. A comparison with a state-of-the-art supervised segmentation approach shows that the novel method is only outperformed if a large amount of labeled training data is available.

## Acknowledgments

This work was supported by the German Research Foundation (DFG) under grant no. ME3737/3-1.

## References

- Aicha BenTaieb and Ghassan Hamarneh. Topology aware fully convolutional networks for histology gland segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'16)*, pages 460–468, 2016.
- Agisilaos Chartsias, Thomas Joyce, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Adversarial image synthesis for unpaired multi-modal cardiac data. In *Proceedings of the International MICCAI Workshop Simulation and Synthesis in Medical Imaging (SASHIMI'17)*, pages 3–13, 2017.
- Michael Gadermayr, Ann-Kathrin Dombrowski, Barbara Mara Klinkhammer, Peter Boor, and Dorit Merhof. CNN cascades for segmenting whole slide images of the kidney. *CoRR*, <https://arxiv.org/abs/1708.00251>, 2017.

- Michael Gadermayr, Vitus Appel, Barbara M. Klinkhammer, Peter Boor, and Dorit Merhof. Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'18)*, pages 165–173, 2018a.
- Michael Gadermayr, Dennis Eschweiler, Barbara Mara Klinkhammer, Peter Boor, and Dorit Merhof. Gradual domain adaptation for segmenting whole slide images showing pathological variability. In *International Conference on Image and Signal Processing (ICISP'18)*, Springer LNCS, pages 461–469, 2018b.
- N. Herve, A. Servais, E. Thervet, J.-C. Olivo-Marin, and V. Meas-Yedid. Statistical color texture descriptors for histological images analysis. In *Proceedings of ISBI'11*, pages 724–727, 2011.
- Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, and Joel H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the International Conference on Computer Vision (CVPR'16)*, 2016.
- Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *CoRR*, <https://arxiv.org/abs/1802.07934>, 2018. URL <http://arxiv.org/abs/1802.07934>.
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017.
- Alexander J. Ratner, Henry R. Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Re. Learning to compose domain-specific transformations for data augmentation. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS'17)*, 2017.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV'16)*, 2016.
- Tsuyoshi Kato, Raissa Relator, Hayliang Ngouv, Yoshihiro Hirohashi, Osamu Takaki, Tetsuhiro Kakimoto, and Kinya Okada. Segmental HOG: new descriptor for glomerulus detection in kidney microscopy image. *BMC Bioinformatics*, 16(1), 2015.
- Mateusz Kozí Nski, Loïc Simon, and Frédéric Jurie. An Adversarial Regularisation for Semi-Supervised Training of Structured Output Neural Networks. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS'17)*, 2017.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Aided Interventions (MICCAI'15)*, pages 234–241, 2015.
- Mira Valkonen, Kimmo Kartasalo, Kaisa Liimatainen, Matti Nykter, Leena Latonen, and Pekka Ruusuvuori. Dual structured convolutional neural network with feature augmentation for quantitative characterization of tissue histology. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*, 2017.

Mitko Veta, Paul J. van Diest, and Josien P. W. Pluim. Cutting out the middleman: Measuring nuclear area in histopathology slides without segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'16)*, pages 632–639, 2016.

Jelmer M. Wolterink, Anna M. Dinkla, Mark H. F. Savenije, Peter R. Seevinck, Cornelis A. T. van den Berg, and Ivana Išgum. Deep MR to CT synthesis using unpaired data. In *Proceedings of the International MICCAI Workshop Simulation and Synthesis in Medical Imaging (SASHIMI'17)*, pages 14–23, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the International Conference on Computer Vision (ICCV'17)*, 2017.

# Transfer Learning by Adaptive Merging of Multiple Models

**Robin Geyer**

**Luca Corinzia**

**Viktor Wegmayr**

*ETH Zurich, Institute for Machine Learning, Zurich, Switzerland*

GEYERR@STUDENT.ETHZ.CH

LUCA.CORINZIA@INF.ETHZ.CH

VWEGMAYR@INF.ETHZ.CH

## Abstract

Transfer learning has been an important ingredient of state-of-the-art deep learning models. In particular, it has significant impact when little data is available for the target task, such as in many medical imaging applications. Typically, transfer learning means pre-training the target model on a related task which has sufficient data available. However, often pre-trained models from several related tasks are available, and it would be desirable to transfer their combined knowledge by automatic weighting and merging. For this reason, we propose T-IMM (Transfer Incremental Mode Matching), a method to leverage several pre-trained models, which extends the concept of Incremental Mode Matching from lifelong learning to the transfer learning setting. Our method introduces layer wise mixing ratios, which are learned automatically and fuse multiple pre-trained models before fine-tuning on the new task. We demonstrate the efficacy of our method by the example of brain tumor segmentation in MRI (BRATS 2018 Challange). We show that fusing weights according to our framework, merging two models trained on general brain parcellation can greatly enhance the final model performance for small training sets when compared to standard transfer methods or state-of the art initialization. We further demonstrate that the benefit remains even when training on the entire Brats 2018 data set (255 patients).

**Keywords:** Transfer Learning, Lifelong Learning, Segmentation, Brain, MRI

## 1. Introduction

Machine learning, especially deep learning, has produced impressive results in supervised learning tasks, given that large and densely annotated training data is available (LeCun et al., 2015; Wainberg et al., 2018). However, the generalization performance of deep learning models deteriorates quickly when training data becomes scarce. This condition is one of the reasons that have prevented the extensive use of deep learning models in applications which require expensive annotation, as is often the case in health care.

Transfer learning (TL) (Pan et al., 2010) is a common approach in machine learning to mitigate the lack of target data. It is based on the intuition that humans can learn new tasks quickly even without many examples, because they can rely on previous, similar experiences. Similarly, TL pre-trains a model on a task, which is similar to the target task, but has sufficient training data available. More specifically, the weights of the model are adjusted to minimize the loss of the first learning task, before they are used as initialization for the target task, as shown in fig. 1b. Besides improving generalization, TL also offers a way to share information without sharing sensitive data, because only the model parameters are revealed to the community of interest. Again, this advantage is particularly evident in medical applications, which often exhibit a co-existence of many privately

maintained models. It can be beneficial for the development of new applications to have access to such prior models, but to date it is not clear how to merge knowledge from multiple models at once.

In order to address this question we propose T-IMM (Transfer-Incremental Mode Matching), an algorithm for transfer learning with multiple prior models. The concept of IMM appears in context of life-long learning (Lee et al., 2017). It differs from transfer learning as its original purpose is not better initialization, but the sequential merging of models, which still retains good performance on all the prior tasks (fig. 1c). Our work provides a useful re-interpretation of IMM for transfer learning. Moreover, our extension T-IMM enables automatic, and *adaptive* merging of multiple models, depicted in fig. 1d. By the example of brain tumor segmentation in MR images, we demonstrate that T-IMM provides a better initialization than common IMM, which represents the corner case of uniform model merging.<sup>1</sup>

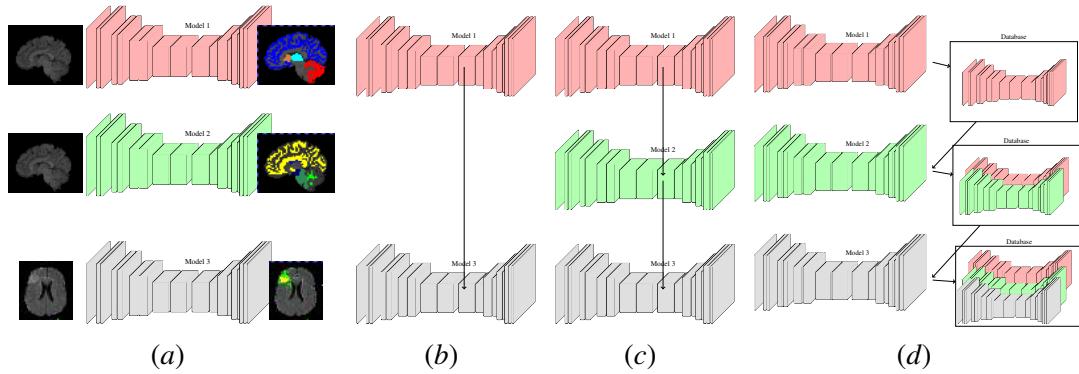


Figure 1: T-IMM framework compared to standard transfer learning approaches. (a) red+green: prior tasks, grey: target task (b) Common transfer learning (c) Sequential IMM (d) T-IMM.

## 2. Related Work

**Transfer Learning** TL encompasses methods that discover shared parameters between prior tasks and a target task (Pan et al., 2010). More specifically, TL improves learning of a target task in three ways (Tommasi et al., 2010): (i) better initial performance, (ii) steeper performance growth, (iii) higher performance at the end of training. Moreover, TL is an important part of many state-of-the-art methods in image classification and segmentation. In these cases, TL is mainly performed by reusing the filter parameters of convolutional neural networks (CNN) such as in the work of (Oquab et al., 2014). They use a CNN pre-trained on ImageNet to compute mid-level image representations for object classification in PASCAL VOC images (Everingham et al., 2012), leading to significantly improved results. To this date, the top scoring submissions to the PASCAL VOC challenge continue to use TL, e.g (Chen et al., 2018) pre-trained on the Coco-data set or (Iglovikov and Shvets, 2018) pre-trained on ImageNet. Despite these success stories, little research has been done on leveraging knowledge from multiple models for a new task. Some work is based on ensemble methods (Gao

1. All the code is available at <https://github.com/cyrusgeyer/TIMM.git>

et al., 2008), which is problematic when the number of available sources is large. A different direction of merging multiple models relies on the particular choice of the SVM model (Tommasi et al., 2010).

**Lifelong learning** Lifelong Learning (LL) describes the scenario when new tasks arrive sequentially, and should be incorporated into the current model one at a time. In contrast to the TL setting, in LL we require to maintain high performance over prior tasks, too. The reason is, when tasks are learned sequentially, performance typically decreases significantly on earlier tasks. This effect is called catastrophic forgetting (Goodfellow et al., 2013), but it is irrelevant for TL, because we usually only care about performance on the target task. Recent developments in LL such as Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2016) and Learning Without Forgetting (LwF) (Li and Hoiem, 2016)) attempt to overcome catastrophic forgetting by regularization of the target loss function. Incremental Moment Matching (IMM) (Lee et al., 2017) does not change the target loss function, but instead provides a parameter merging scheme for a pair of prior models. Specifically, IMM approximates the posterior distribution of parameters for every prior task as a Gaussian with diagonal co-variance, and then computes the parameter distribution for the new task as the best approximation of the prior mixture of Gaussians.

### 3. The T-IMM Method

#### 3.1. Adaptively fusing parameters

Let us consider  $T$  different but related tasks. Moreover, we assume  $T$  models that share the same architecture, and they are trained incrementally on the  $T$  tasks. Incremental training means that the parameters of model  $i$  are used as initialization for model  $i + 1$ . This procedure is in fact necessary, otherwise it would be impossible to maintain a correspondence between parameters. The parameter set  $\Phi$  of each model is partitioned into two sets of parameters, i.e.  $\Phi = \mathcal{P} \cup \mathcal{S}$ ,  $\mathcal{P} \cap \mathcal{S} = \emptyset$ . The first set  $\mathcal{P}$  contains all parameters used for fusion, e.g. convolution filters. The second set  $\mathcal{S}$  contains all task-specific parameters, e.g. batch normalization parameters or the weights of the top-level layers. Following the work of (Lee et al., 2017), we approximate the parameter co-variance matrix of each model with the diagonal of the inverse empirical fisher information matrix. Ignoring off-diagonal entries in the fisher information matrix is critical for our approach, because it allows simple splitting into parameter subsets. Moreover, we can even introduce multiple IMM-mixing ratios for different subsets of  $\mathcal{P}$ , e.g. one ratio for each layer. This enables layer-specific, adaptive merging of models. More formally, let the parameter subset  $\mathcal{P}_t$  of each task  $t$  be composed of  $N$  parameter vectors:  $\mathcal{P}_t = \{\boldsymbol{\theta}_i^t\}_{i=1}^N$ .

We also introduce the set  $\mathcal{F}_t$  which holds the Fisher information matrices for each parameter vector in  $\mathcal{P}_t$ , i.e.  $\mathcal{F}_t = \{\mathbf{F}_i^t\}_{i=1}^N$ .

Lastly, set  $\mathcal{A}$  holds mixing coefficients according to which the parameters in  $\mathcal{P}$  will be fused:

$$\mathcal{A} = \{\boldsymbol{\alpha}^t\}_{t \in \{1 \dots T\}} \quad \text{where } \boldsymbol{\alpha}^t = (\alpha_1^t \dots \alpha_N^t)^T \quad \text{and} \quad \sum_t \boldsymbol{\alpha}^t = \mathbf{1}^N \quad (1)$$

The fused parameters of the new model  $T + 1$  are given by (Lee et al., 2017)

$$\begin{aligned} \mathcal{P}_{T+1} &= \mathcal{P}_{T+1}(\mathcal{A} | \{\mathcal{P}_t, \mathcal{F}_t\}_{t=1}^T) = \{\boldsymbol{\theta}_i^{T+1}\}_{i \in \{1 \dots N\}} \\ \text{where } \boldsymbol{\theta}_i^{T+1} &= \left( \sum_t \boldsymbol{\alpha}_i^t \mathbf{F}_i^t \right)^{-1} \sum_t \boldsymbol{\alpha}_i^t \mathbf{F}_i^t \boldsymbol{\theta}_i^t \end{aligned} \quad (2)$$

### 3.2. Equally weighted IMM Transfer

Without any further information, the choice of the mixing coefficients is arbitrary. The common IMM method assumes equally weighted merging, hence it sets  $\alpha_i^t = 1/T$ , for all tasks  $t$  and layers  $i$ . After the task-specific parameters  $\mathcal{S}_{T+1}$  are randomly initialized, the entire model  $\Phi_{T+1}$  is fine-tuned on task  $T + 1$ .

### 3.3. T-IMM

To achieve the best possible performance on the new task, we desire to find a better non-uniform mixing. However, it is clearly impractical to search the space of all possible  $\mathcal{A}$  manually, in particular if  $T$  is large. For this reason, the Transfer-IMM (T-IMM) method splits transfer learning into two stages: a short adaption stage and an extensive fusing stage.

**Adaption Stage** In the adaption stage, we aim to learn the mixing coefficients that are best suited for transferring knowledge. We start by randomly initializing the task-specific layers in  $\mathcal{S}_{T+1}$ . The merged parameters  $\mathcal{P}_{T+1}$  are initialized according to eq. (2), as a function of  $\mathcal{A}$ . The adaption stage optimization can be formalized as:

$$\underset{\mathcal{A}, \mathcal{S}}{\text{minimize}} \quad Loss_{T+1}(\mathcal{A}, \mathcal{S} | \mathcal{P}_{T+1}(\cdot)) \quad \text{subject to} \quad \sum_t^T \alpha^t = \mathbf{1}^N \text{ and } \alpha^t \succeq 0, \forall t \quad (3)$$

The constraints on the mixing coefficients can be enforced reparametrizing the mixing ratios. We introduce a set of new unconstrained variables  $\{\delta^t\}_{t \in \{1, \dots, T\}}$  and using the sigmoid activation function  $\sigma$  we can write the mixing ratios as:

$$\alpha_i^t = \frac{\sigma(\delta_i^t)}{\sum_{j=1}^T \sigma(\delta_j^t)} \quad \text{with} \quad \delta_j^t \in \mathbb{R}$$

The adaption stage terminates once the loss converges, and returns a set of mixing coefficients adapted to the new task.

**Fine tuning stage** After determination of the mixing ratios  $\mathcal{A}$  according to eq. (3), all parameters  $\Phi_{T+1} = (\mathcal{P}_{T+1}, \mathcal{S}_{T+1})$  are fine-tuned until convergence on a validation set. We also reuse  $\mathcal{S}_{T+1}$  from the adaption stage as initialization for the fine-tuning stage. More formally:

$$\underset{\mathcal{P}, \mathcal{S}}{\text{minimize}} \quad Loss_{T+1}(\mathcal{P}, \mathcal{S})$$

The T-IMM method is depicted in fig. 2.

## 4. Experiments and Results

### 4.1. FCNN architecture

**Medical image segmentation** Fully convolutional neural networks (fCNN) are state of the art in medical image segmentation (2D and 3D). For instance, in segmentation of brain MRI, both BRATS and MRBrainS challenges are lead by fCNN-approaches. This is also the case for interactive segmentation, where 2D- and 3D-fCNNs define the state-of-the-art (Wang et al., 2017a,b). Therefore,

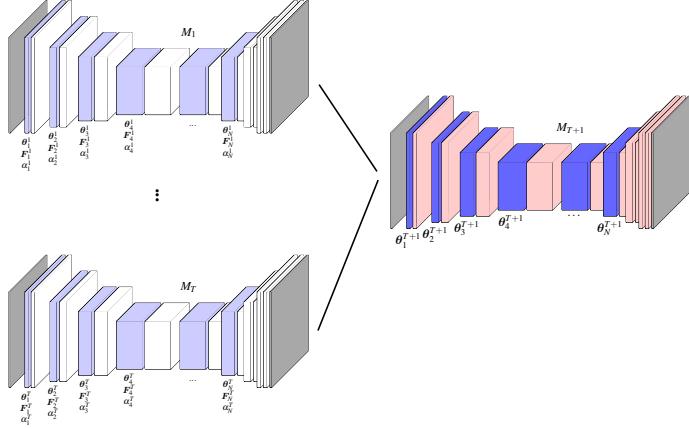


Figure 2: T-IMM framework:  $T$  models trained on  $T$  different tasks are merged as initialization for task  $T + 1$ . All the parameters in the light-blue layers are in the  $\mathcal{P}$  set. The parameters of the new model  $\{\theta_i^{T+1}\}$  in the dark blue layers are merged according to eq. (2). The parameters of the red layers (parameters in  $\mathcal{S}$ ) account for task specific layers (i.e. batch-norm, instance normalization layers, or the last classification layers) and hence are randomly initialized.

we also choose this type of neural network for our demonstration of T-IMM in medical image segmentation. Specifically, we use the fCNN-architecture proposed by (Isensee et al., 2017). This network is inspired by the U-net (Ronneberger et al., 2015), and is comprised of 30 3D-convolutional layers, and 27 instance normalization layers. The task specific parameters  $\mathcal{S}$  contain the three top-level segmentation layers, and all instance normalization layers. The remaining convolutional filter parameters comprise the set  $\mathcal{P}$  of transferred parameters.

#### 4.2. Data and Tasks

In our experiments we perform three different tasks of brain tissue and brain tumor segmentation. Two tasks concern the parcellation of different brain regions, while the third task is about brain tumor segmentation. While the first two tasks are learned incrementally, the third task is learned with different initializations: random, transfer from either of the two prior tasks, initialization with IMM of model 1+2, and initialization with T-IMM of model 1+2.

**Task 1 & Task 2** The data used for the first and second task are brain MR images from the Human Connectome Project. The data set holds a total of 58 brain images with T1+T2 weighting. For each brain, 14 different classes were annotated by (Karani et al., 2018), using the FreeSurfer software: (1) cerebellum gray matter (2) Cerebral gray matter, Cortex and Accumbens, (3) Thalamus, (4) Amygdala and Choroid Plexus, (5) Caudate, (6) Pallidum, (7) Celebrospinal Fluid, (8) Cerebellum white matter, (9) Cerebral white matter, (10) Hippocampus, (11) Ventricel, (12) Putamen, (13) Ventral DC ,(14) Brainstem. The data is split into three non-overlapping groups of size 23,23 and 12 respectively. Task 1 is defined as segmenting labels 1 through 7 on the first split of 23 brains. Task 2 is defined as segmenting labels 8 through 14 on the second split of 23 brains. The third split of 12 brains is the test set both tasks can be evaluated on.

Table 1: Description of the different tasks and datasets.

	Total	Training	Validation	Labels
Task 1	23	18	5	1:7
Task 2	23	18	5	8:14
4 %	10	8	2	ET, TC, WT
Task 3	8 %	20	16	ET, TC, WT
	100 %	255	215	ET, TC, WT
	Total			
Test set Task 1 & Task 2			12	1:14
Test set Task 3			40	ET, TC, WT
Online-Val set Task 3			66	ET, TC, WT

**Task 3** The third task is brain tumor segmentation as defined by the BRATS-2018 challenge. The data set holds 255 patients. For each patient we have four different modalities (T1,T1w,T2,Flair) and an expert’s annotation of the enhancing tumor (ET), the tumor core (TC) and the whole tumor (WT). Furthermore, we are provided an online validation set (Online-Val) of 66 non-annotated patients. We evaluate our framework for different portions of the total data: using 4 %, 8% and 100% of the Brats data-set. 40 brains are used as a test set to evaluate the individual experiments on. For the 100 %-experiments, theses 40 brains are used as the validation set and the online, non-annotated set of 66 patients is used as a test set. This is in order to make our experimental results comparable to each other but also to state of the art benchmarks. An outline of the different tasks and datasets is available in table 1.

**Data preprocessing** We conduct very simple data preprocessing. For data used in Task 3 (Brats), ANTS N4-Bias field correction is conducted. Furthermore, all data is histogram-normalized to filter out irrelevant differences.

### 4.3. Experiments

**Testing the Framework** We start by training an fCNN model on Task 1 (M1). After convergence we use the parameters of M1 to initialize M2, which is then trained on task 2 until convergence. Having trained M1 and M2, the main experiments are conducted. For each data portion (4%, 8% and 100%) the following five initialization methods for task 3 are tested: Xavier random initialization (referred to as ‘No Transfer’), Parameter Transfer from Model 1, Parameter Transfer from Model 2, Parameter Transfer using IMM and Parameter Transfer using T-IMM.

**Understanding T-IMM** In order to better understand T-IMM, we further conduct the following three experiments (only for the 8% portion due to computational reasons): parameter transfer from a model that was trained on all HCP-data and on all labels 1:14, parameter transfer from a model that was trained on all HCP-data but only on labels 1:7 and parameter transfer from a model that was trained on all HCP-data but only on labels 8:14. All these are then compared to 8% T-IMM.

**Transfer Learning and catastrophic forgetting** In a last experiment we evaluate how much of tasks 1 & 2 is remembered by the different initialization models. For this, the task specific sets of parameters  $\mathcal{S}_1, \mathcal{S}_2$  of M1 and M2 are used in combination with the parameter sets  $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_{IMM}, \mathcal{P}_{T-IMM}$ .

#### 4.4. Evaluation

The metric we chose to asses an fCNN’s performance is the Dice Coefficient (DC) averaged over all relevant labels reached on the test set examples. table 2 holds our main results for 4%, 8% and 100% of Brats data. We are dealing with paired-samples, i.e. for each patient in the test set we evaluate the DC difference between T-IMM and all other methods. The differences are visualized in fig. 4. The numbers reported on the 100%-scenario are the feedback of submissions of the 66-unannotated examples to the BRATS validation leader-board. Table 3 shows the length of the different training stages of T-IMM and the final validation score reached at the end of the adaption stage. The distribution of the mixing ratios after the adaption stage is displayed in the appendix Table 4 evaluates how well the models used for transfer remember tasks 1 and 2. This is also visualized in fig. 3.

Table 2: Mean Dice Coefficient of ET,TC and WT for different transfer scenarios and different portions of Brats data

Transfer:	No	Model 1	Model 2	IMM	T-IMM	All HCP all labels	All HCP labels 1:7	All HCP labels 8:14
4%	0.30	0.39	0.52	0.55	<b>0.58</b>	-	-	-
8%	0.55	0.60	0.61	0.63	<b>0.65</b>	0.60	0.63	0.63
100%	0.79	0.81	0.81	0.81	<b>0.82</b>	-	-	-

Table 3: Epochs needed and validation dice score reached for adaption stage and fine-tuning stage

	Adaption Stage		Fine Tuning Stage	
	Epochs	Val-score	Epochs	Val-score
4 %	16	0.44	88	0.72
8 %	20	0.40	178	0.68
100 %	10	0.67	129	0.77

Table 4: Mean dice coefficient when the models used for initialization are evaluated on the test set of Task 1 & Task 2

	Labels 1:7	Labels 8:14
M1	0.89	0.00
M2	0.01	0.89
IMM	0.38	0.42
T-IMM	0.50	0.39

#### 5. Discussion

We can assert multiple things from the results in table 2 and fig. 4. We see the fact confirmed, that the benefit of transfer learning grows with smaller training sets. This was shown in several studies before. We also see that M1 and M2 are not equally well suited for parameter transfer. Especially for the 4% and 8%-scenario, model 2 clearly brings more advantage than model 1. However, initializing with model 1 still outperforms no transfer. T-IMM manages to solve the dilemma of having to choose a priori which model to transfer knowledge from. The experiments for 8% of Brats data show that initializing training with T-IMM even outperforms initialization with a model that was trained on all tasks and all data that T-IMM is able to fuse from. table 4 shows, that both IMM

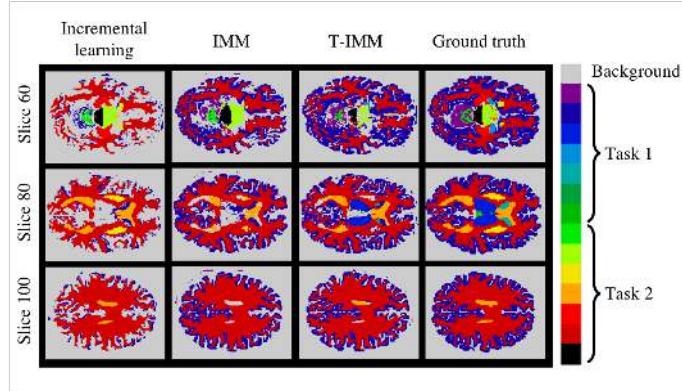


Figure 3: Catastrophic forgetting of task 1, which is learned before task 2. Each task consists of the parcellation of seven different brain regions (colors), and background (grey). **First column:** Prediction after fine-tuning on task 2 using model from task 1 as initialization. The model forgot to predict the blue classes from task 1. **Second column:** Prediction of model obtained with IMM between models trained on task 1 and 2. Clearly reduced forgetting of task 1. **Third column:** Prediction of model obtained after the *adaption stage* of T-IMM. Even though the model was fused to perform a third task, it remembers tasks 1+2 very well. **Fourth column:** Ground truth class labels.

and T-IMM do overcome catastrophic forgetting to a certain extend and manage to remember task 1 and task 2 (even though with lower performance). For T-IMM this is especially interesting, as the model underwent the adaption stage, were it trains to suit task 3. This reassures the assumption that indeed, feature representations from both models/tasks are reused by T-IMM. We were able to show that fusing different CNN for parameter transfer using T-IMM can give a decisive advantage over settling for a singe transfer source, especially when training data is sparse. We further show that the advantage shrinks but remains significant even for large data sets.

## References

- Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs/1802.02611, 2018. URL <http://arxiv.org/abs/1802.02611>.
- M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>, 2012.
- Jing Gao, Wei Fan, Jing Jiang, and Jiawei Han. Knowledge transfer via multiple model local structure mapping. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 283–291. ACM, 2008.

- Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.
- Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation. *CoRR*, abs/1801.05746, 2018. URL <http://arxiv.org/abs/1801.05746>.
- Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. *2017 International MICCAI BraTS Challenge*, 2017. URL [https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI\\_BraTS/MICCAI\\_BraTS\\_2017\\_proceedings\\_shortPapers.pdf](https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf).
- Neerav Karani, Krishna Chaitanya, Christian Baumgartner, and Ender Konukoglu. A Lifelong Learning Approach to Brain MR Segmentation Across Scanners and Protocols. *arXiv e-prints*, art. arXiv:1805.10170, May 2018.
- James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016. URL <http://arxiv.org/abs/1612.00796>.
- Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 521:436–444, 2015.
- Sang-Woo Lee, Jin-Hwa Kim, JungWoo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. *CoRR*, abs/1703.08475, 2017. URL <http://arxiv.org/abs/1703.08475>.
- Zhizhong Li and Derek Hoiem. Learning without forgetting. *CoRR*, abs/1606.09282, 2016. URL <http://arxiv.org/abs/1606.09282>.
- Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.
- Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL <http://arxiv.org/abs/1505.04597>.
- Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *Proceedings of IEEE Computer Vision and Pattern Recognition Conference*, 2010.
- Michael Wainberg, Daniele Merico, Andrew Delong, and Brendan J. Frey. Deep learning in biomedicine. *Nature Biotechnology*, 36:829–838, 2018.

Guotai Wang, Wenqi Li, Maria A. Zuluaga, Rosalind Pratt, Premal A. Patel, Michael Aertsen, Tom Doel, Anna L. David, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Interactive medical image segmentation using deep learning with image-specific fine-tuning. *CoRR*, abs/1710.04043, 2017a. URL <http://arxiv.org/abs/1710.04043>.

Guotai Wang, Maria A. Zuluaga, Wenqi Li, Rosalind Pratt, Premal A. Patel, Michael Aertsen, Tom Doel, Anna L. David, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Deepigeos: A deep interactive geodesic framework for medical image segmentation. *CoRR*, abs/1707.00652, 2017b. URL <http://arxiv.org/abs/1707.00652>.

## Appendix A. Subject-level comparison of T-IMM vs. other initializations

In this section we show the sample-wise performance of T-IMM method compared to other methods on the Brats set. The samples are sorted in ascending order according to the performance of T-IMM.

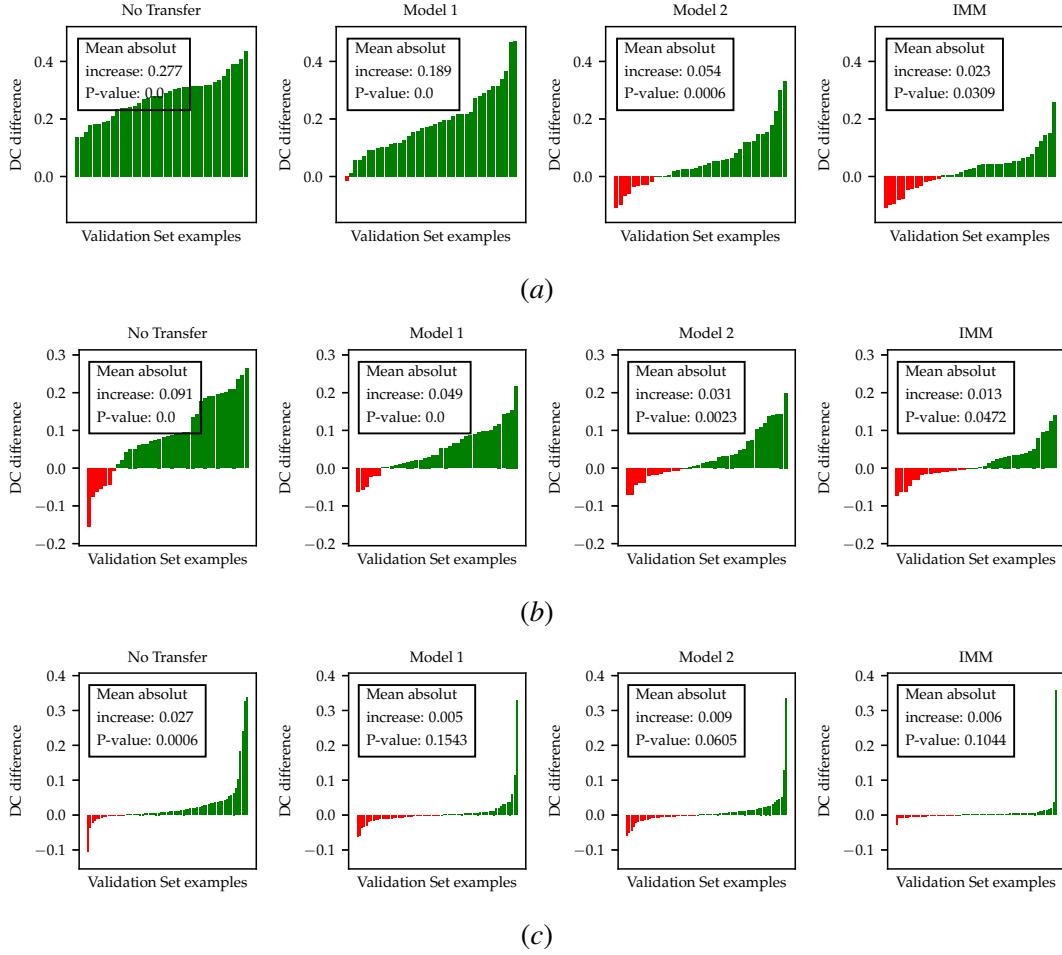


Figure 4: Absolute Dice coefficient increase of T-IMM compared to other transfer methods for training an fCNN on (a): 4 % of Brats data, (b) 8% of Brats data and on (c) 100% of Brats data (the 100% is evaluated on the online validation set)

## Appendix B. Mixing ratios after adaption stage of T-IMM

In this section we show the distribution of mixing ratios after the adaption stage of T-IMM was completed and before the fine tuning stage started.

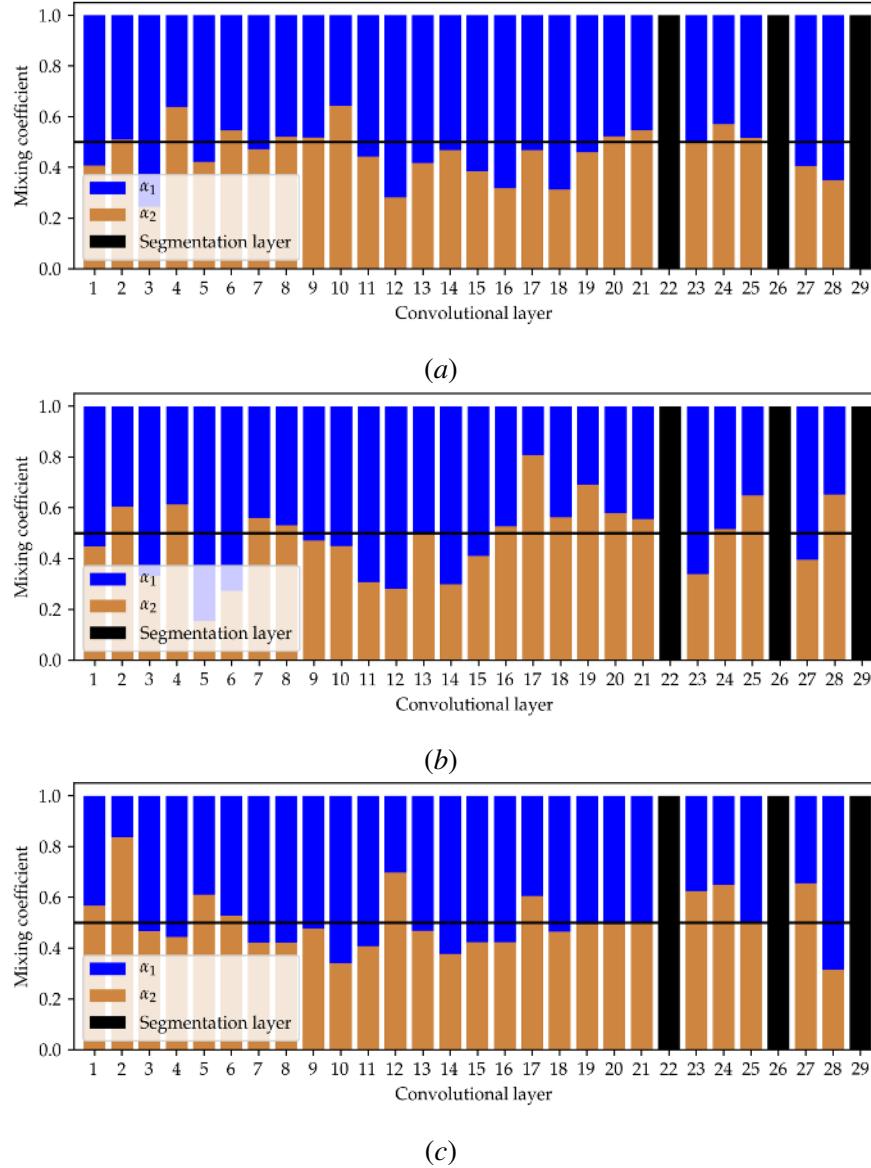


Figure 5: Mixing ratios for each convolutional layer in the set  $\mathcal{P}$  of transferable weights, after the adaption stage of T-IMM was completed. Experiments of (a): 4 % of Brats data, (b) 8% of Brats data and on (c) 100% of Brats data (the 100% is evaluated on the online validation set)

# Assessing Knee OA Severity with CNN attention-based end-to-end architectures

**Marc Górriz**<sup>1</sup>

**Joseph Antony**<sup>2</sup>

**Kevin McGuinness**<sup>2</sup>

**Xavier Giró-i-Nieto**<sup>1</sup>

**Noel E. O’Connor**<sup>2</sup>

ALGAYON2@GMAIL.COM

JOSEPH.ANTONY@DCU.IE

KEVIN.MCGUINNESS@DCU.IE

XAVIER.GIRO@UPC.EDU

NOEL.OCONNOR@DCU.IE

<sup>1</sup> Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia/Spain

<sup>2</sup> Insight Centre for Data Analytics, Dublin City University (DCU), Dublin, Ireland

## Abstract

This work proposes a novel end-to-end convolutional neural network (CNN) architecture to automatically quantify the severity of knee osteoarthritis (OA) using X-Ray images, which incorporates trainable attention modules acting as unsupervised fine-grained detectors of the region of interest (ROI). The proposed attention modules can be applied at different levels and scales across any CNN pipeline helping the network to learn relevant attention patterns over the most informative parts of the image at different resolutions. We test the proposed attention mechanism on existing state-of-the-art CNN architectures as our base models, achieving promising results on the benchmark knee OA datasets from the osteoarthritis initiative (OAI) and multicenter osteoarthritis study (MOST). All code from our experiments will be publicly available on the github repository: <https://github.com/marc-gorriz/KneeOA-CNNAttention>

**Keywords:** Convolutional Neural Network, End-to-end Architecture, Attention Algorithms, Medical Imaging, Knee Osteoarthritis.

## 1. Introduction

Knee osteoarthritis (OA) is the most common articular disease and a leading cause of chronic disability (Heidari, 2011), and mainly affects the elderly, obese, and those with a sedentary lifestyle. Degenerative processes of the articular cartilage as a result of excessive load on the joint, and the aging process, contributes to the natural breakdown of joint cartilage with joint space narrowing (JSN) and osteophytes (Li et al., 2013). Knee OA causes excruciating pain and often leads to joint arthroplasty in its severe stages. An early diagnosis is crucial for clinical treatment to be effective in curtailing progression and mitigating future disability (Oka et al., 2008) (Shamir et al., 2009a). Despite the introduction of several imaging modalities such as MRI, OCT, and ultrasound for augmented OA diagnosis, X-Ray is still the method of choice in diagnosing knee OA, although clinical evidence also contributes.

Previous work has approached the challenge of automatically assessing knee OA severity as an image classification problem (Shamir et al., 2009a) (Shamir et al., 2008) (Yoo et al., 2016) using the Kellgren and Lawrence (KL) grading (Kellgren and Lawrence, 1957). KL grading quantifies the degree of degeneration on a five-point scale (0 to 4): KL-0 (no OA changes), KL-1 (doubtful),

KL-2 (early OA changes), KL-3 (moderate), and KL-4 (end-stage). Assessment is based on JSN, presence of osteophytes, sclerosis, and bone deformity. Most methods in the literature use a two step process to automatically quantify knee OA severity: 1) localization of knee joints; and 2) quantification of severity. Separate models for knee joint localization, either using hand-crafted features ([Shamir et al., 2008](#)), ([Yoo et al., 2016](#)) or CNNs ([Antony et al., 2016](#)) are not always highly accurate, affecting the subsequent quantification accuracy and adding extra complexity to the training process.

To overcome this problem, this work proposes a novel end-to-end architecture incorporating trainable attention modules that act as unsupervised fine-grained ROI detectors, which automatically localize knee joints without a separate localization step. The proposed attention modules can be applied at different levels and scales across an arbitrary CNN pipeline. This helps the network to learn attention patterns over the most informative parts of the image at different resolutions, achieving improvements in the quantification performance.

## 2. Related work

Much of the literature has proposed image classification-based solutions to assess knee OA severity using radiography-based semi-quantitative scoring systems, like KL gradings, which are based on the study of anatomical features such as variations in joint space width or osteophytes formation ([Shamir et al., 2009b](#)) ([Shamir et al., 2009a](#)) ([Shamir et al., 2008](#)). Shamir et al. ([Shamir et al., 2008](#)) proposed WND-CHARM: a multi purpose medical image classifier to automatically assess knee OA severity in radiographs using a set of features based on polynomial decompositions, contrast, pixel statistics, textures, and features from image transforms. Recently, Yoo et al. ([Yoo et al., 2016](#)) proposed a self-assessment scoring system associating risk factors and radiographic knee OA features using multivariable logistic regression models, additionally using an Artificial Neural Network (ANN) to improve the overall scoring performance. Shamir et. al. ([Shamir et al., 2009a](#)) proposed template matching to automatically detect knee joints from X-ray images. This method is slow to compute for large datasets and gives poor detection performance. Antony et al. ([Antony et al., 2016](#)) introduced an SVM-based approach for automatically detecting the knee joints. Later, Antony et al. ([Antony et al., 2017](#)) proposed an FCN-based approach to improve the localization of the knee joints. Although more accurate, the aspect ratio chosen for the extracted knee joints affects the overall quantification.

Recently, the emergence of deep learning has enabled the development of new intelligent diagnostics based on computer vision. CNNs outperform many state-of-the-art methods based on hand-crafted features in tasks such as image classification ([Krizhevsky et al., 2012](#)), retrieval ([Babenko et al., 2014](#)) and object detection ([Lawrence et al., 1997](#)) ([Wei et al., 2011](#)). Antony et al. ([Antony et al., 2016](#)) showed that the off-the-shelf CNNs such as the VGG 16-layer network ([Simonyan and Zisserman, 2014](#)), the VGG-M-128 network ([Chatfield et al., 2014](#)), and the BVLC reference CaffeNet ([Jia et al., 2014](#)) ([Karayev et al., 2013](#)) pre-trained on ImageNet LSVRC dataset ([Russakovsky et al., 2015](#)) can be fine-tuned for classifying knee OA images through transfer learning. They argued that it is appropriate to assess knee OA severity using continuous metrics like mean-squared error together with binary or multi-class classification losses, showing that predicting the continuous grades through regression reduces the error and improves overall quantification. They proposed a novel pipeline ([Antony et al., 2017](#)) to automatically quantify knee OA severity using a FCN for localization and a CNN jointly trained for classification and regression. The work consolidates the

state-of-the-art baseline for the application of CNNs in the field, opening a range of research lines for further improvements. Tiulpin et al. ([Tiulpin et al., 2018](#)) presented a new computer-aided diagnosis method based on using deep Siamese CNNs, which are originally designed to learn a similarity metric between pairs of images. However, rather than comparing image pairs, the authors extend this idea to similarity in knee x-ray images (with 2 symmetric knee joints). Splitting the images at the central position and feeding both knee joints into a separate CNN branch allows the network to learn identical weights for both branches. They outperform the previous approaches by achieving an average multi-class testing accuracy score of 66.71 % on the entire OAI dataset, despite also needing a localization step to focus the network branches on the knee joint areas.

This work mainly focuses on designing an end-to-end architecture with attention mechanisms. There are similar methods reported in the literature. Xiao et al. ([Xiao et al., 2015](#)) propose a pipeline to apply visual attention to deep neural networks by integrating and combining attention models to train domain-specific nets. In another approach, Liu et al. ([Liu et al., 2016](#)) introduce a reinforcement learning framework based on fully convolutional attention networks (FCAN) to optimally select local discriminative regions adaptive to different fine-grained domains. The proposed weakly-supervised reinforcement method combined with a fully-convolutional architecture achieves fast convergence without requiring expensive annotation. Recently, Jetley et al. ([Jetley et al., 2018](#)) introduce an end-to-end-trainable attention module for CNN architectures built for image classification. The module takes as input the 2D feature vector maps, which forms the intermediate representations of the input image at different stages in the CNN pipeline, and outputs a matrix of scores for each map. They redesign standard architectures to classify the input image using only a weighted combination of local features, forcing the network to learn relevant attention patterns.

### 3. Method

This section describes the proposed methods, detailing the design, implementation, and training of the attention modules. Several strategies are investigated to integrate the attention mechanism into standard CNN architectures, proposing experimental approaches to classify the knee images.

#### 3.1. Trainable Attention Module for CNNs

The selected attention module is inspired by the work of Kevin Mader ([Mader, 2018](#)), in which a trainable attention mechanism is designed for a pretrained VGG-16 network to predict bone age from hand X-Ray images. Figure 1 illustrates this idea.

Given an input volume  $D^l$  from a convolutional layer  $l$  with  $N$  feature maps, several  $1 \times 1$  convolutional layers are stacked to extract spatial features. The output is then passed to a  $1 \times 1$  locally connected layer ([Chen et al., 2015](#)) (convolution with unshared weights) with sigmoidal activation to give an attention mask  $A^l$ . The original feature maps are element-wise multiplied by the attention mask, generating a new convolutional volume  $\tilde{D}^l$  accentuating informative areas. A spatial dimensionality reduction is performed by applying global average pooling (GAP) on the masked volume, generating a  $N$ -dimensional feature vector  $F^l$ , which is then normalized by the average value of the attention mask. Additionally, a *softmax* layer can be applied to yield a  $C$ -dimensional vector with the output class probabilities  $\{p_1, p_2, \dots, p_C\}$ .

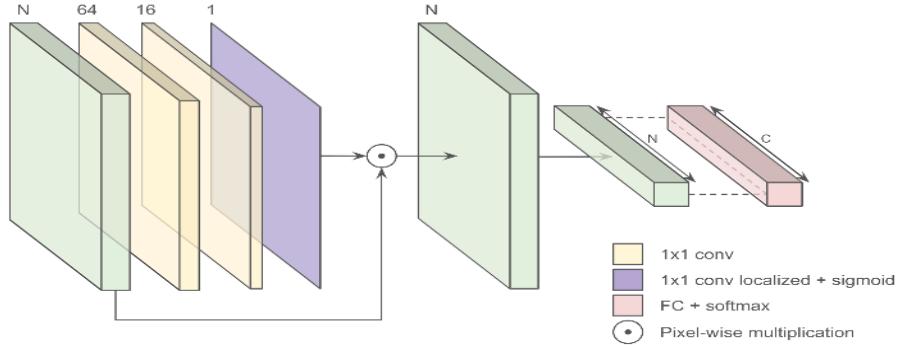


Figure 1: Attention module scheme.

### 3.2. Module Integration to CNN Pipeline

Standard CNN architectures typically stack several convolutional layers with occasional pooling operations that reduces the spatial dimension and increase the receptive field. Therefore, the degree of abstraction of the attention modules is closely related to their location in the CNN, focusing on more global details as depth increases. We define the concept of an attention branch as the location of an attention module in a specific convolutional block, applying a *softmax* operation at the top to produce independent class probabilities based on the KL scores. Each attention branch will be seen as a new model by itself that could be trained end-to-end. Figure 2 shows a sample architecture integrating the attention modules to the VGG-16 pipeline. Fixing an input size of  $320 \times 224$  pixels, as fixed in Section 3.4, we build the branches  $\text{att}\{i\}_{i=0..2}$  taking as input volumes the feature maps belonging to the pooling layers after the convolutional blocks 3, 4, and 5. Following the methodology in Section 3.3, a combinational module is applied to fuse the local features from all the branches into a global feature vector and then generate the KL grades probability distribution by applying a *softmax* layer at the top.

### 3.3. Combining Multiple Attention Branches

Several strategies are investigated to merge features from multiple branches with the aim to combine attention patterns at different resolutions. Our first strategy is performing early fusion of features from different branches. Each attention module generates a  $N$ -dimensional feature vector  $F^l$  with the average values of the  $N$  masked feature maps conforming the input convolutional volume. A channel-wise concatenation is applied to fuse all the branches, generating a new vector  $F_A = [F^{l_1} F^{l_2} \dots F^{l_B}]$  with  $F_A \in R^{1x(\sum N_b)}$ , being  $B$  the total of attention branches and  $N_b$  the dimension of  $F^{l_b}$ . In addition, a fully connected layer is added at the top to perform early fusion of the concatenated features, while a *softmax* operation is applied to generate the  $C$ -dimensional output class probabilities. As shown above, the complexity of the attention modules correlates with their location in the CNN pipeline, which biases the convergence behavior. This can be critical for a combined model that attempts to train modules with different convergence rates at the same time: deeper branches quickly overfit while waiting for the convergence of the slower ones. In contrast,

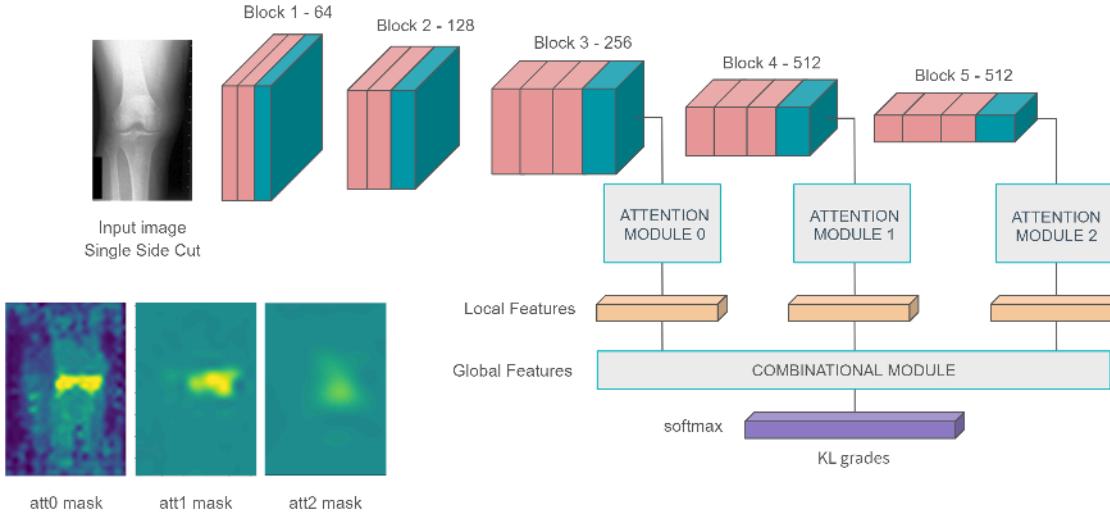


Figure 2: Sample architecture integrating the attention modules in the VGG-16 pipeline.

by reducing the overall learning time, the shallower branches with more complex modules may decrease their performance due to under training.

Our next strategy is to simplify the multiple branch learning process. We propose the use of multi-loss training, which aims to improve the learning efficiency and prediction performance by learning multiple objectives from a shared representation. Each attention branch makes separate predictions via a *softmax* to generate their class probabilities and we linearly combine the individual categorical cross-entropies  $\mathcal{L}_b$  into a global loss:  $\mathcal{L} = \sum_b w_b \mathcal{L}_b$ , with  $\mathcal{L}_b = -\sum_c y_c^{(b)} \log p_c^{(b)}$ . This allows to control the rate of convergence by weighting the contribution of each branch, assigning low weights for those branches with faster convergence to reduce their influence at the initial stages of training and attenuate updates in shallower attention modules. There are previous approaches that propose the use of multi-loss training to address different machine learning tasks such as dense prediction (Kokkinos, 2017), scene understanding (Eigen and Fergus, 2015), natural language processing (Collobert and Weston, 2008) or speech recognition (Huang et al., 2013). However, the model performance is extremely sensitive to the weight selection  $w_b$ , that needs an expensive and time-consuming hyper-parametrization process.

Several multi-branch combinations were tested by applying multidimensional cross-validation to find the optimum branch locations and multi-loss weights. We used a 2D grid search, validating the *att0* and *att1* loss weights between a range of 0.5 to 1 with a step size of 0.1, and using the validation loss as monitor. The best performance was achieved with *att0*, *att1* weights  $w_0 = 1$ ,  $w_1 = 0.8$ , slightly reducing the contribution of the deeper attention modules and decreasing their overfitting tendency while the shallower branches are still learning.

### 3.4. Public Knee OA Datasets

The data used for this work are bilateral PA fixed flexion knee X-ray images. The datasets are from the Osteoarthritis Initiative (OAI) and Multicenter Osteoarthritis Study (MOST) in UCSF, being standard public datasets widely used in knee OA studies. The baseline cohort of the OAI dataset

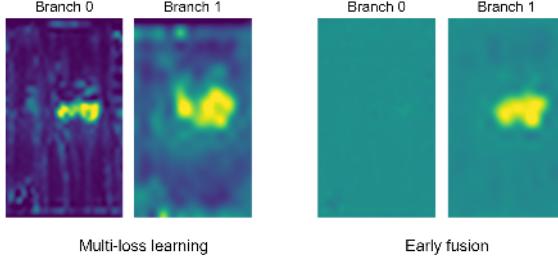


Figure 3: Comparison between merging solutions in the VGG-16 pipeline, visualizing the generated masks in the attention branches  $att_0$  and  $att_1$  and observing a large improvement in the shallower modules using multi-loss.

contains MRI and X-ray images of 4,476 participants. From this entire cohort, we selected 4,446 X-ray images based on the availability of KL grades for both knees as per the assessments by Boston University X-ray reading center (BU). The MOST dataset includes lateral knee radiograph assessments of 3,026 participants. From this, 2,920 radiographs are selected based on the availability of KL grades for both knees as per baseline assessments. As a pre-processing step, all the X-ray images are manually split in the middle, generating two vertical sections from the left and right sides, resizing them to a fixed mean size of  $320 \times 224$  pixels by keeping the average aspect ratio. Histogram equalization is performed for intensity level normalization, and eventually data augmentation is applied by performing horizontal right-left flips to generate more training data. The training, validation, and test sets were split based on the KL grades distribution. A 70-30 train-test split was used and 10% of the training data was kept for validation.

### 3.5. Training

All models were trained from scratch using categorical cross-entropy with the ground truth KL grades. Regarding multi-branch training, all target data were duplicated for each attention branch. We used Adam ([Kingma and Ba, 2014](#)) with a batch size of 64,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , an initial learning rate of  $10^{-5}$  scaled by 0.1 every 2 epochs without improvement in validation loss, and early stopping after 3 epochs without improvement.

## 4. Results

Although the attention mechanism can be integrated to any CNN pipeline, not all the architectures are well-suited for assessing knee OA severity. We explored several architectures in the literature including state-of-the-art models from previous works of Antony et al. ([Antony et al., 2017](#)) and more complex architectures such as ResNet-50 ([He et al., 2016](#)), and we analyzed their performance. We found that the same level of abstraction in  $att_1$  and  $att_2$  for Antony et al models can be achieved in shallower branches  $att_0$ ,  $att_1$  for deeper architectures, implying the best branch location depends on model complexity. After testing different branch combinations, the best performing locations presented in Table 1, and detailed in the Tables 3, 4, 5, specifying the output resolution of their convolutional blocks and the location of the attention branches. From the evaluation of

multi-loss models, since every single attention branch produces an independent prediction, the top performing one is used at test time. A more sophisticated ensemble approach was considered but not included in this paper. This approach involves averaging the pre-activation outputs (i.e. values before the softmax) of each of the model branches and then passing the result through a softmax to enforce a valid probability distribution over the classes. This idea is often effective in test time data augmentation and ensemble methods and may improve performance over the single best model referred here. As Table 1 shows, the VGG-16 attention branch *att0* with multi-loss learning achieved the best overall classification accuracy (64.3%).

Table 1: Evaluation overview for different CNN pipelines.

Models	Antony Clsf.	Antony Extended	ResNet-50	VGG-16
<i>att0</i>	41.76%	40%	59.3%	62.2%
<i>att1</i>	41.9%	53.26%	58.67%	61.7%
<i>att2</i>	44.53%	53.83%	56.47%	58.8%
Early fusion	52.5%	56.17%	59%	63%
Multi-Loss	<i>att1</i> : 43.9% <i>att2</i> : 46.6%	<i>att1</i> : 55.61% <i>att2</i> : 55.68%	<i>att0</i> : <b>60%</b> <i>att1</i> : 56.88%	<i>att0</i> : <b>64.3%</b> <i>att1</i> : 63.2%

We also compared the attention mechanism with related knee OA classification-based solutions in the literature. First, we retrained the Antony et. al models with the same training data from the previous experiments, applying the FCN introduced by the authors to address the knee joints extraction (Antony et al., 2017). The results (Table 2) show that the attention-based models with end-to-end architectures clearly outperform the state-of-the-art frameworks. We further compared our results to human level accuracy using the radiologic reliability readings from the OAI (Klara et al., 2016). Although the data used to compute the reliability grading does not match our test set, we followed the methodology of previous works in the literature (Tiulpin et al., 2018), with the aim to dispose of a panoramic view of the current gold standard for diagnosing OA involving human performance. Cohen’s *kappa* coefficient (Cohen, 1960) was used to evaluate the agreement between non-clinician readers and experienced radiologists by classifying *N* items with *C* mutually exclusive categories. Considering the following grading:  $\kappa < 0.20$  slight agreement, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, and 0.81-1.0 almost perfect agreement, their inter-reader reliability for the KL scores was moderate to substantial, with  $\kappa$  values between 0.5 and 0.8. In the case of automatic assessments, by considering a CNN model as a non-clinician X-Ray reader, we can apply the  $\kappa$  coefficient to evaluate the inter-reader reliability between its predictions and the corresponding ground truth annotations, provided by experienced radiologists. As Table 2 shows, the *att0* branch of the VGG-16 trained by multi-loss together with the *att1* branch ( $w_1 = 0.8$ ), improves the reliability of related works with a substantial  $\kappa = 0.63$  agreement, reaching the margins of human accuracy.

## 5. Conclusions

This work proposed a novel end-to-end architecture that incorporates trainable attention modules acting as unsupervised fine-grained ROI detectors. The proposed attention modules can be applied

Table 2: Comparative with related frameworks

	<b>Test Acc.</b>	<b>Test Loss</b>	<b>Kappa</b>	<b># parameters</b>
Antony Clsf	59%	1.13	0.42	~ 7.51 M (FCN: ~ 212.7 K)
Antony Joint Clsf & Reg	62.29%	0.89	0.48	~ 5.77 M (FCN: ~ 212.7 K)
VGG-16: Multi-loss ( <i>att0</i> )	<b>64.3%</b>	0.817	<b>0.63</b>	~ 7.7 M

at different levels and scales across the CNN pipeline, helping the network to learn relevant attention patterns over the most informative parts of the image at different resolutions. The results obtained for the public knee OA datasets OAI and MOST were satisfactory despite having a considerable scope for further improvement.

The proposed attention mechanism can be easily integrated to any convolutional neural network architecture, being adaptable to any input convolutional volume. However, after exploring different off-the-shelf base models for classification with different complexities, we observed that the best performance is achieved in those models with a balanced ratio between the complexity of the overall architecture and the depth of the convolutional volumes, avoiding overfitting while getting abstraction in the local features used to train the attention modules. On the other hand, we propose the use of multi-loss training to manage the training of multiple attention branches with different velocities of convergence at the same time, boosting the overall performance by fusing attention features with different levels of abstraction. The best performance was achieved by slightly reducing the contribution of the deepest attention branches, improving then the precision of the shallower attention masks and reaching the effectiveness of related approaches with a test accuracy of 64.3% and Kappa agreement of 0.63. Although our method does not surpass the state-of-the-art and could be interpreted as challenging to implement, the overall aim was to reduce the training complexity using an end-to-end architecture. As mentioned in Section 4, without an end-to-end design, the models require a localization step to focus the classifier to the knee joint regions of interest. For instance, previous work of Antony et. al. (Antony et al., 2017) needed a manual annotation process for training a FCN to automatically segment the input knee joints. Our approach, in contrast, requires no such annotation of knee joint locations in the training data. Finally, we observed that localizing the knee joints in an unsupervised way can reduce performance by adding noise in the attention masks and thus into the overall process. A more robust attention module can improve the results and have a bigger impact in the future. As future work, it may be interesting to design better base networks for the attention mechanism and then to test new fine-grained methods in the state-of-art, with the aim to improve the performance of the attention modules towards reducing their dependence on the complexity of the base model.

## Acknowledgments

This research was supported by contract SGR1421 by the Catalan AGAUR office. The work has been developed in the framework of project TEC2016-75976-R, funded by the Spanish Ministerio de Economía y Competitividad and the European Regional Development Fund (ERDF). The authors also thank NVIDIA for generous hardware donations.

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant numbers SFI/12/RC/2289 and 15/SIRG/3283.

The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2- 2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. MOST is comprised of four cooperative grants (Felson – AG18820; Torner – AG18832; Lewis – AG18947; and Nevitt – AG19069) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by MOST study investigators. This manuscript was prepared using MOST data and does not necessarily reflect the opinions or views of MOST investigators.

## References

- Joseph Antony, Kevin McGuinness, Noel E O’Connor, and Kieran Moran. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 1195–1200. IEEE, 2016.
- Joseph Antony, Kevin McGuinness, Kieran Moran, and Noel E O’Connor. Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 376–390. Springer, 2017.
- Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *European conference on computer vision*, pages 584–599. Springer, 2014.
- Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.
- Yu-hsin Chen, Ignacio Lopez-Moreno, Tara N Sainath, Mirkó Visontai, Raziel Alvarez, and Carolina Parada. Locally-connected and convolutional neural networks for small footprint speaker recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.
- Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.
- David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

- Behzad Heidari. Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part i. *Caspian journal of internal medicine*, 2(2):205, 2011.
- Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7304–7308. IEEE, 2013.
- Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.
- Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.
- Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013.
- JH Kellgren and JS Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases*, 16(4):494, 1957.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Kristina Klara, Jamie E Collins, Ellen Gurary, Scott A Elman, Derek S Stenquist, Elena Losina, and Jeffrey N Katz. Reliability and accuracy of cross-sectional radiographic assessment of severe knee osteoarthritis: role of training and experience. *The Journal of rheumatology*, pages jrheum–151300, 2016.
- Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, volume 2, page 8, 2017.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997.
- YongPing Li, XiaoChun Wei, JingMing Zhou, and Lei Wei. The age-related changes in cartilage and osteoarthritis. *BioMed research international*, 2013, 2013.
- Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016.
- Kevin Mader. Attention on pretrained-vgg16 for bone age. <https://www.kaggle.com/kmader/attention-on-pretrained-vgg16-for-bone-age>, 2018.

- H Oka, S Muraki, T Akune, A Mabuchi, T Suzuki, H Yoshida, S Yamamoto, K Nakamura, N Yoshimura, and H Kawaguchi. Fully automatic quantification of knee osteoarthritis severity on plain radiographs. *Osteoarthritis and Cartilage*, 16(11):1300–1306, 2008.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- Lior Shamir, Nikita Orlov, D Mark Eckley, Tomasz Macura, Josiah Johnston, and Ilya G Goldberg. Wndchrm—an open source utility for biological image analysis. *Source code for biology and medicine*, 3(1):13, 2008.
- Lior Shamir, Shari M Ling, William Scott, Marc Hochberg, Luigi Ferrucci, and Ilya G Goldberg. Early detection of radiographic knee osteoarthritis using computer-aided analysis. *Osteoarthritis and Cartilage*, 17(10):1307–1312, 2009a.
- Lior Shamir, Shari M Ling, William W Scott Jr, Angelo Bos, Nikita Orlov, Tomasz J Macura, D Mark Eckley, Luigi Ferrucci, and Ilya G Goldberg. Knee x-ray image analysis method for automated detection of osteoarthritis. *IEEE Transactions on Biomedical Engineering*, 56(2):407–415, 2009b.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Aleksei Tiulpin, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Scientific reports*, 8(1):1727, 2018.
- Jun Wei, Heang-Ping Chan, Chuan Zhou, Yi-Ta Wu, Berkman Sahiner, Lubomir M Hadjiiski, Marilyn A Roubidoux, and Mark A Helvie. Computer-aided detection of breast masses: Four-view strategy for screening mammography. *Medical physics*, 38(4):1867–1876, 2011.
- Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015.
- Tae Keun Yoo, Deok Won Kim, Soo Beom Choi, and Jee Soo Park. Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study. *PloS one*, 11(2):e0148724, 2016.

## Appendix A. CNN Architectures, Learning Curves and Visualizations

Table 3: Antony et al. base architecture for classification

<b>Layer</b>	<b>Kernels</b>	<b>Kernel Size</b>	<b>Strides</b>	<b>Output shape</b>
conv1	32	$11 \times 11$	2	$100 \times 150 \times 32$
pool1	-	$3 \times 3$	2	$49 \times 74 \times 32$
conv2	64	$5 \times 5$	1	$49 \times 74 \times 64$
(att0) pool2	-	$3 \times 3$	2	$24 \times 36 \times 64$
conv3	96	$3 \times 3$	1	$24 \times 36 \times 96$
(att1) pool3	-	$3 \times 3$	2	$11 \times 17 \times 96$
conv4	128	$3 \times 3$	1	$11 \times 17 \times 128$
(att2) pool4	-	$3 \times 3$	2	$5 \times 8 \times 128$

Table 4: Antony et al. extended base architecture for classification and regression

<b>Layer</b>	<b>Kernels</b>	<b>Kernel Size</b>	<b>Strides</b>	<b>Output shape</b>
conv1	32	$11 \times 11$	2	$100 \times 150 \times 32$
pool1	-	$3 \times 3$	2	$49 \times 74 \times 32$
conv2-1	64	$3 \times 3$	1	$49 \times 74 \times 64$
conv2-2	64	$3 \times 3$	1	$49 \times 74 \times 64$
(att0) pool2	-	$3 \times 3$	2	$24 \times 36 \times 64$
conv3-1	96	$3 \times 3$	1	$24 \times 36 \times 96$
conv3-2	96	$3 \times 3$	1	$24 \times 36 \times 96$
(att1) pool3	-	$3 \times 3$	2	$11 \times 17 \times 96$
conv4-1	128	$3 \times 3$	1	$11 \times 17 \times 128$
conv4-2	128	$3 \times 3$	1	$11 \times 17 \times 128$
(att2) pool4	-	$3 \times 3$	2	$5 \times 8 \times 128$

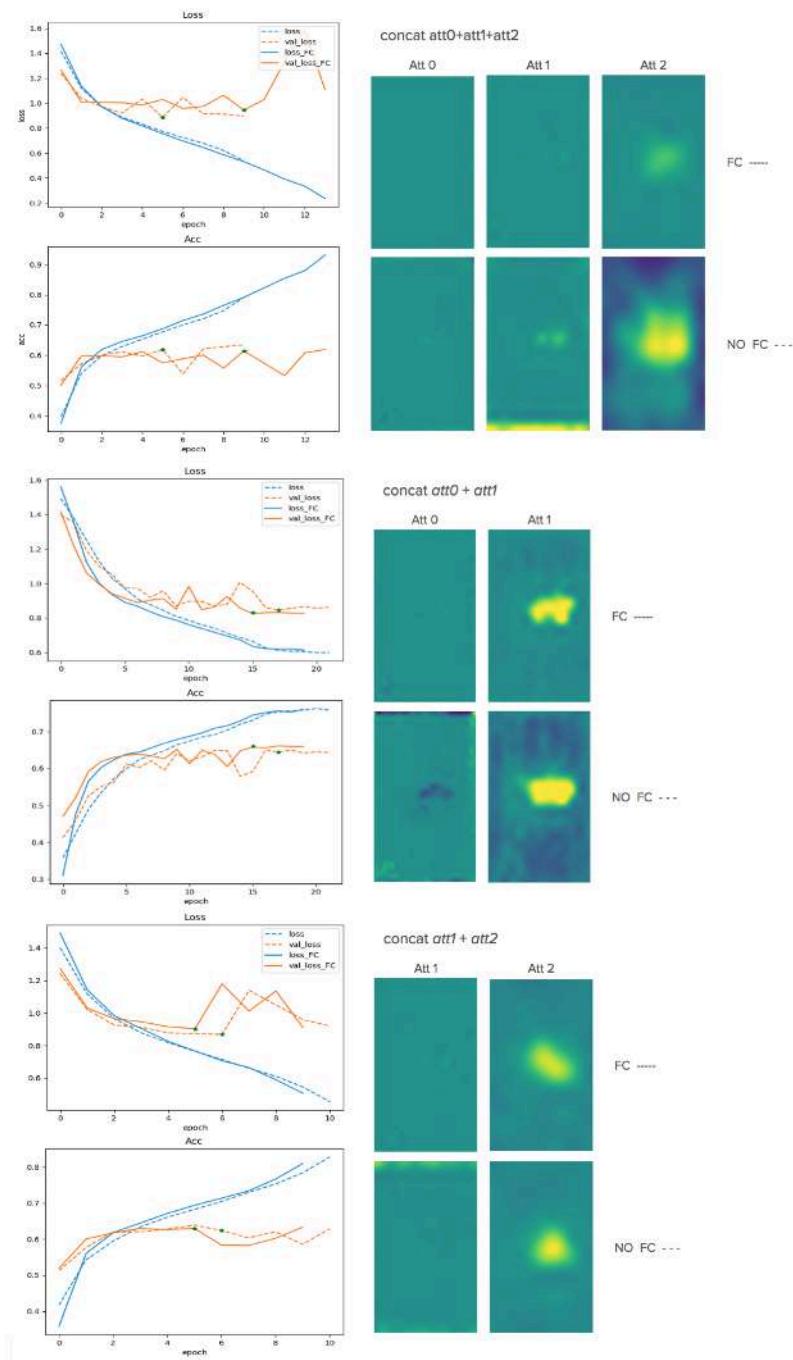


Figure 4: Learning curves and visualization for early fusion experiment in VGG-16.

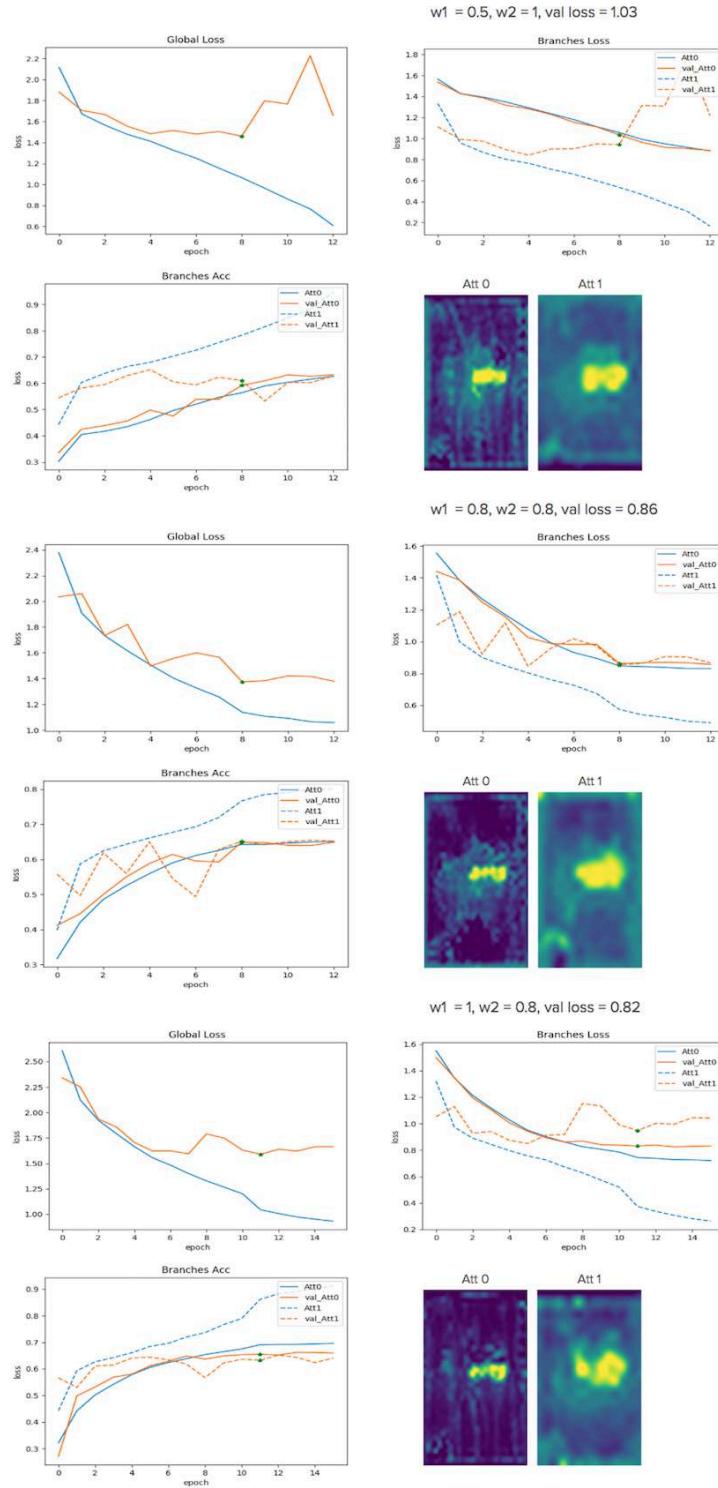


Figure 5: Learning curves and visualization for multi-loss experiment in VGG-16.

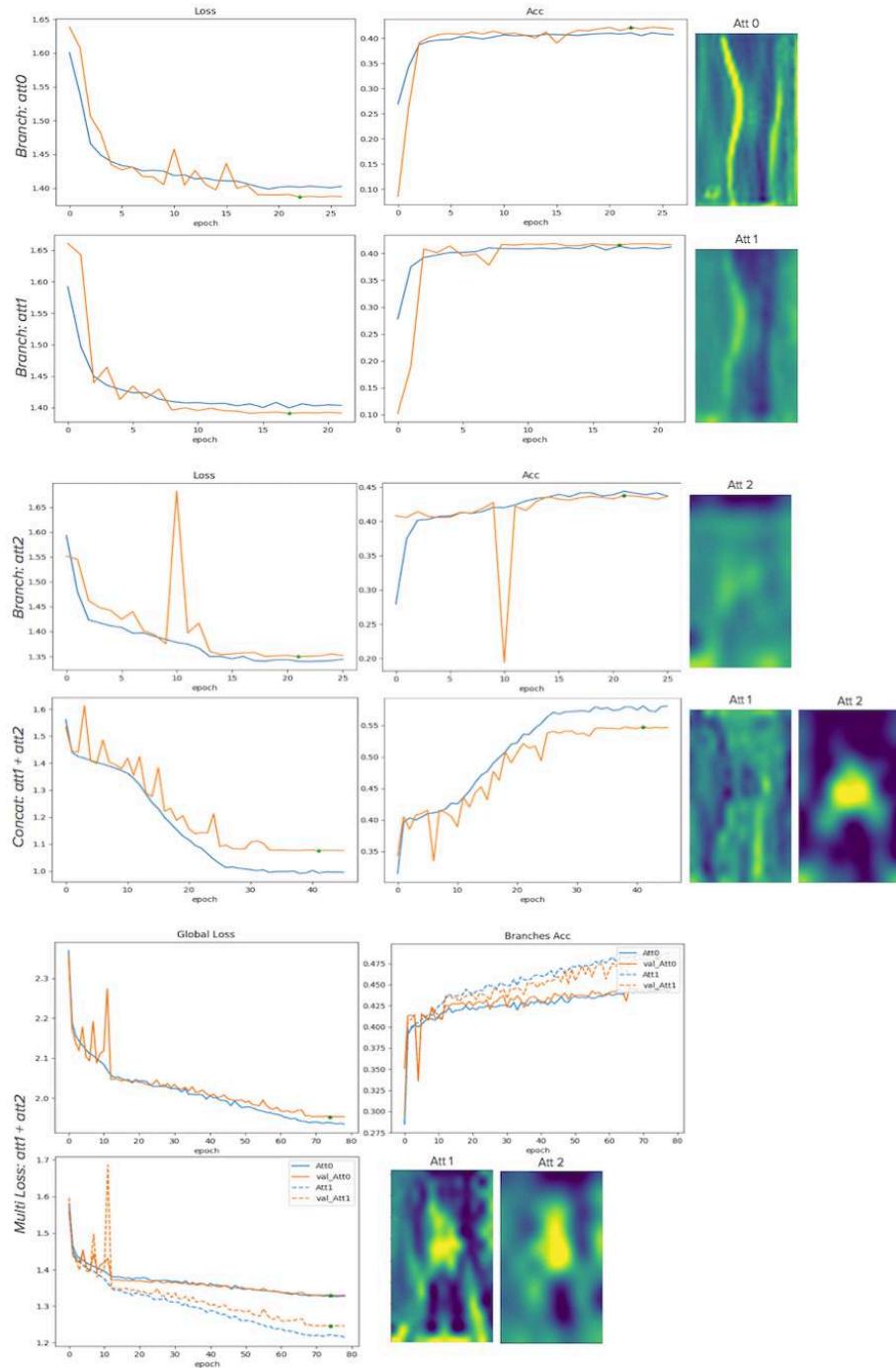


Figure 6: Learning curves and visualization for Antony et al. pipeline for classification.

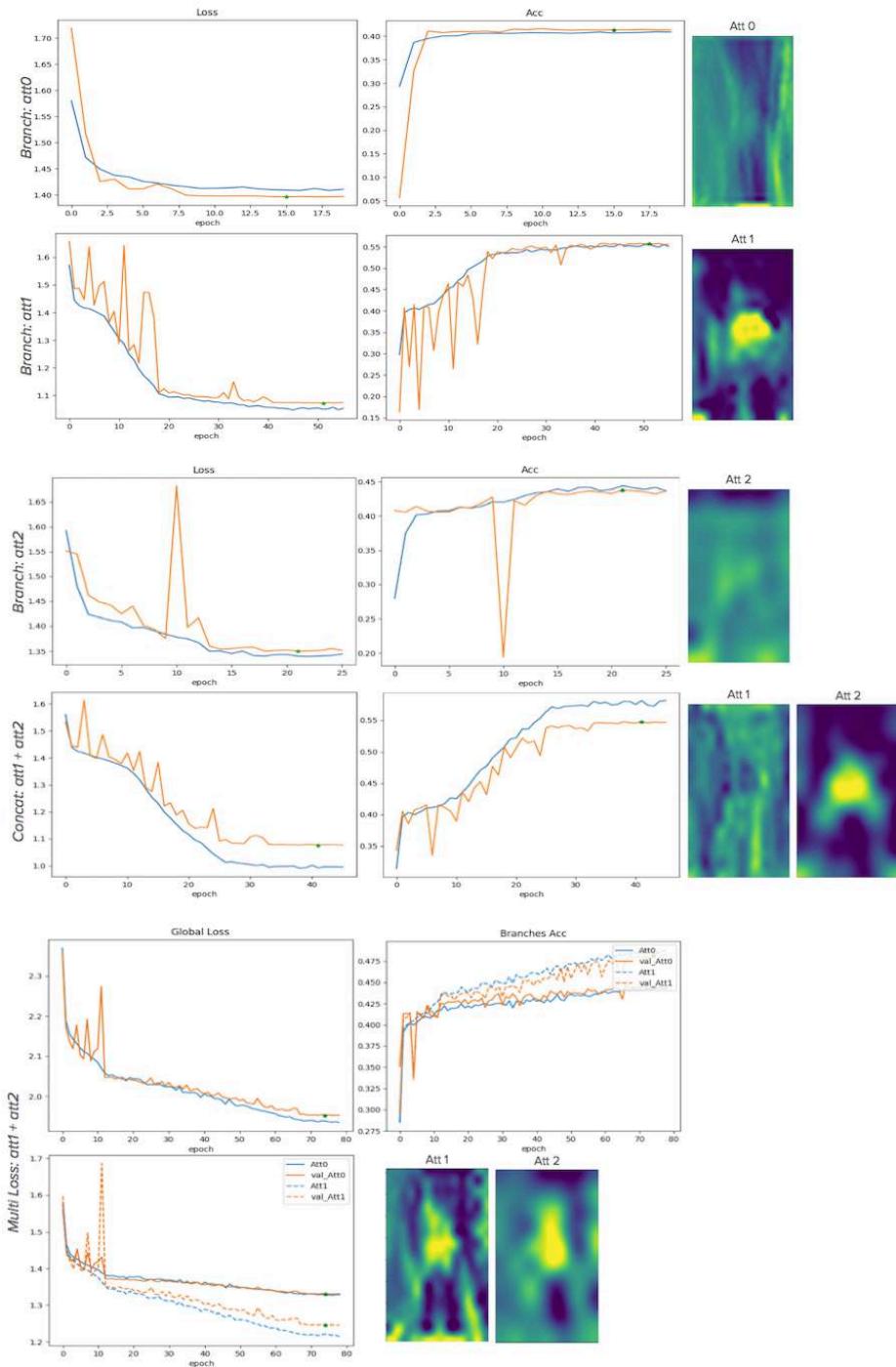


Figure 7: Learning curves and visualization for Antony et al. pipeline for jointly classification and regression.

Table 5: ResNet-50 base architecture for classification

<b>Layer</b>	<b>Kernels</b>	<b>Kernel Size</b>	<b>Strides</b>		<b>Output shape</b>
conv1	64	$7 \times 7$	2		$224 \times 224 \times 64$
maxpool	-	$3 \times 3$	2		$112 \times 112 \times 64$
conv2_*	64	$1 \times 1$	1		
	64	$3 \times 3$	1	( $\times 3$ )	$56 \times 56 \times 256$
	256	$1 \times 1$	1		
(att0) conv3_*	128	$1 \times 1$	2		
	128	$3 \times 3$	1	( $\times 4$ )	$28 \times 28 \times 512$
	512	$1 \times 1$	1		
(att1) conv4_*	256	$1 \times 1$	2		
	256	$3 \times 3$	1	( $\times 6$ )	$14 \times 14 \times 512$
	1024	$1 \times 1$	1		
(att2) conv5_*	512	$1 \times 1$	2		
	512	$3 \times 3$	1	( $\times 3$ )	$7 \times 7 \times 512$
	2048	$1 \times 1$	1		

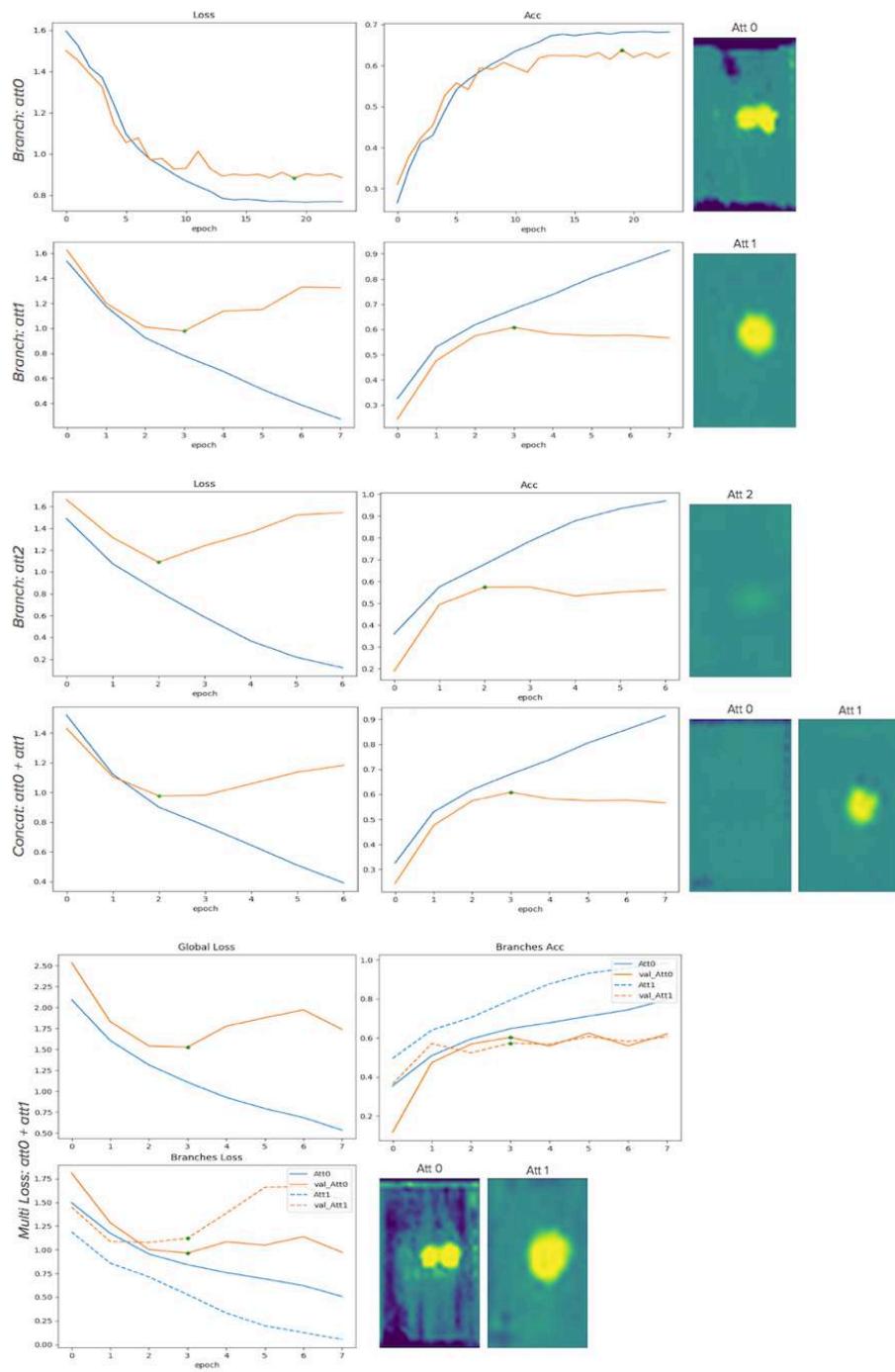


Figure 8: Learning curves and visualization for ResNet-50 pipeline.

# Iterative learning to make the most of unlabeled and quickly obtained labeled data in histology

**Laxmi Gupta<sup>1</sup>**

LAXMI.GUPTA@LFB.RWTH-AACHEN.DE

<sup>1</sup> Institute of Imaging & Computer Vision, RWTH Aachen University, Aachen, Germany

**Barbara Mara Klinkhammer<sup>2</sup>**

BKLINKHAMMER@UKAACHEN.DE

**Peter Boor<sup>2</sup>**

PBOOR@UKAACHEN.DE

<sup>2</sup> Institute of Pathology, University Hospital Aachen, RWTH Aachen University, Aachen, Germany

**Dorit Merhof<sup>\*1</sup>**

DORIT.MERHOF@LFB.RWTH-AACHEN.DE

**Michael Gadermayr<sup>\*1,3</sup>**

MICHAEL.GADERMAYR@FH-SALZBURG.AC.AT

<sup>3</sup> Salzburg University of Applied Sciences, Salzburg, Austria

## Abstract

Due to the increasing availability of digital whole slide scanners, the importance of image analysis in the field of digital pathology increased significantly. A major challenge and an equally big opportunity for analyses in this field is given by the wide range of tasks and different histological stains. Although sufficient image data is often available for training, the requirement for corresponding expert annotations inhibits clinical deployment. Thus, there is an urgent need for methods which can be effectively trained with or adapted to a small amount of labeled training data. Here, we propose a method to find an optimum trade-off between (low) annotation effort and (high) segmentation accuracy. For this purpose, we propose an approach based on a weakly supervised and an unsupervised learning stage relying on few roughly labeled samples and many unlabeled samples. Although the idea of weakly annotated data is not new, we firstly investigate the applicability to digital pathology in a state-of-the-art machine learning setting.

**Keywords:** Digital pathology, convolutional neural networks, kidney, segmentation, weakly supervised.

## 1. Introduction

One of the most significant limiting factors in histological image analysis via machine learning is the need for data annotated by experts (Komura and Ishikawa, 2018). What limits the easy availability of such data is the fact that each whole slide image (WSI) typically has a resolution of about one giga pixels and annotations are required at pixel level. Furthermore, histological images exhibit several dimensions for variability, such as differences in stainings (Gadermayr et al., 2018), etc., which limit generalization of methods developed for a certain data set. Because of these limitations, it is difficult and expensive to obtain annotations.

Researchers have proposed several solutions to deal with these challenges. One option (a) is to optimize the labeling procedure to obtain maximum label data with a fixed manual effort (Komura and Ishikawa, 2018). Another option (b) is to perform effective augmentation of the available labeled data set (Ronneberger et al., 2015). If training data for a similar task is available, transfer

---

\* Contributed equally

learning (c) can be applied in order to adjust a pre-trained model (Gadermayr et al., 2018). The pre-trained model is trained on sufficient data on the related task with few labeled samples, unlabeled samples or a mix of both.

These approaches either require sufficient training data for a similar task (c), or they do not focus on making use of unlabeled data (a,b). In the field of digital pathology, however, typically huge amounts of data are routinely captured despite the fact that exhaustive labeling of all data is mostly unfeasible or at least highly uneconomical. In order to make use of the great potential of unlabeled data, here we focus on semi-supervised learning (Khoreva et al., 2017).

[Khoreva et al. \(2017\)](#) evaluate a scenario involving semantic and instance segmentation of 'easy-to-segment' objects using only coarse annotations for training. They generate training labels from bounding boxes (BBs) of the ROIs and train a convolutional neural network iteratively, employing modification cues at each iteration. The cues are based on the labeled BBs and prior information about the objects to be segmented. Their experiments establish that the model benefits from a recursive training returning object shapes significantly better than the input BBs. With their setting, the authors show that segmentation accuracies similar to fully-supervised approaches can be reached using only BBs as annotations.

The first part of our work is inspired by this idea. Their work is based on Pascal VOC12 and COCO datasets, which in comparison to histological images, are easy to segment. In the former, the objects, typically cats and dogs, can be separated from their background class quite distinctly based on color or gradient information. However, histological data consists of rather textured information, with cells, the typical regions of interest (ROIs), resembling their background class quite closely (see Fig. 1).

### Contribution:

In this paper, we propose and analyze a pipeline to find an optimum trade-off between segmentation accuracy and manual annotation effort for effectively training segmentation models. For that purpose, we develop a two-stage semi-supervised approach that incorporates a weakly supervised and an unsupervised training method. Inspired by [Khoreva et al. \(2017\)](#), in the first stage we train a fully convolutional network in a weakly supervised way utilizing only a limited amount of quickly obtained rough annotations. In the second stage, we exploit the fact that in digital pathology often large amounts of unlabeled data are available. This data is further used for unsupervised optimization of the model based on specifically developed constraints by incorporating statistical prior knowledge. Here, we investigate the applicability of the method in kidney pathology. Specifically, we segment glomeruli on WSIs of mouse kidney ([Kato et al., 2015](#); [Herve et al., 2011](#)) (see Fig. 1). We pose the question for the most effective stage as well as for the best combination of the two stages.

## 2. Methods

### Segmentation Model:

For segmentation, we adapt the method proposed in [Gadermayr et al. \(2019\)](#) (details in Section 3) which is a state-of-the-art approach based on the U-Net architecture ([Ronneberger et al., 2015](#)) yielding high accuracies for the same task. As suggested in [Gadermayr et al. \(2019\)](#), we extract training patches randomly from all over the kidney to incorporate data uniformly from the tissue section. We also maintain the ratio of 2:1 for patches with true positives (TPs) and random patches,

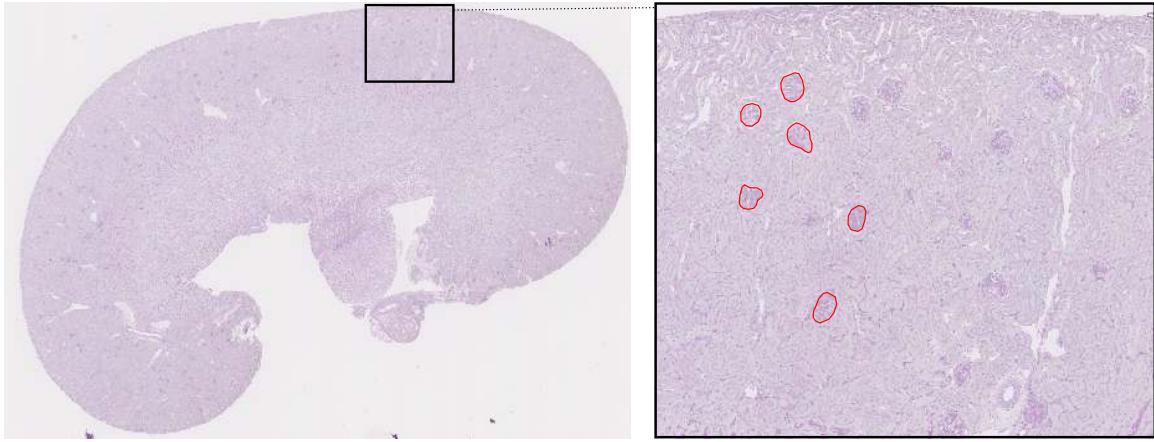


Figure 1: Example WSI of a mouse kidney (left) as well as a magnification of a small patch showing the glomeruli (some are marked in red color, right).

as suggested in the paper for all our experiments. We set the batch size to five, image patch size to 492 and trained with the Adam optimizer on a tensorflow framework. The number of epochs was optimized for each experiment individually (details in Section 2.1).

### Proposed Pipeline:

As motivated earlier in this paper, we develop a two stage approach to make the best use of the available weakly labeled (in Stage 1) and unlabeled (in Stage 2) data. The proposed pipeline is outlined in Fig. 2.

As shown in the figure, the training of the network in Stage 1 begins with BBs. The BB of a single object is defined by the minimum rectangle enclosing the object (see Fig. 3). After the first training round, the training images (IS-train1) are segmented using the model thus obtained. These segmentations are then post processed with the help of 'Cues 1' (details in Section 2.1) before using them as the new training labels for the successive iteration. This process is continued for N-iterations.

For Stage 2, the basic iterative training approach remains the same, except for two important differences. Firstly, the initial training labels are generated using the final model trained in Stage 1. This model is applied to the unseen IS-train2 image data set. Secondly, the set of constraints used to modify intermediate results is different (Cues 2) due to the unsupervised setting (we cannot make use of any ground truth (GT) information in Stage 2, for details see Section 2.2).

Ultimately the resulting models from both stages are tested on the test data set IS-test. Further details about the stages are given in the following subsections.

### 2.1. Stage 1 – Weakly supervised learning

To inspect the applicability of the method to histological images, we adapt the method proposed by Khoreva et al. (2017) to histological data. We apply a very similar approach as proposed in their paper, that involves iterative training and integration of prior information (Cues 1) about the data in the procedure. We initiate the training with BBs as labels. The obtained segmentations, which are

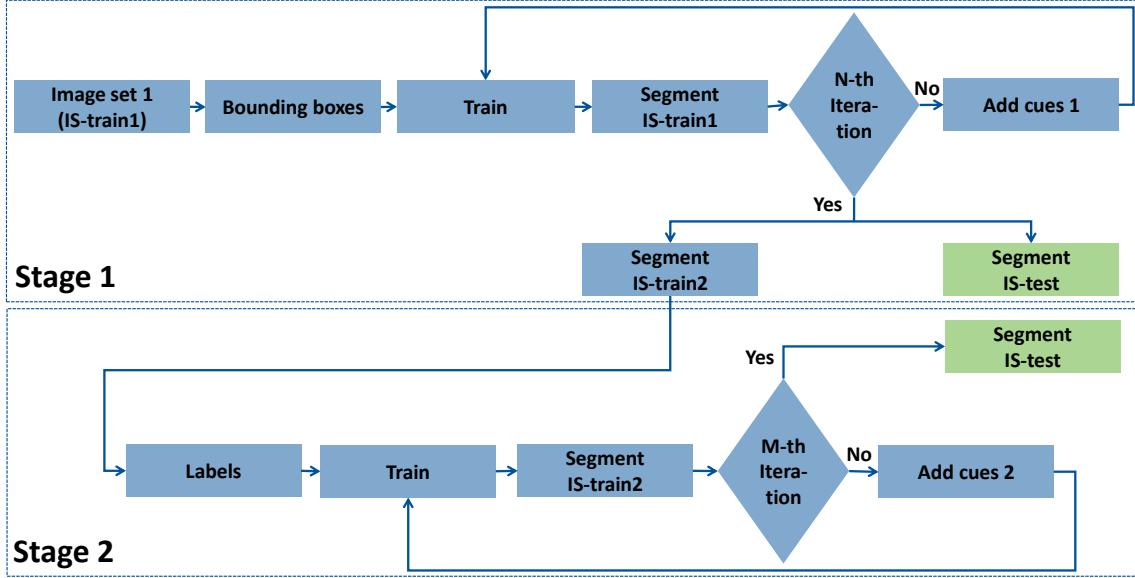


Figure 2: Schematic of the proposed pipeline: Stage 1 shows the weakly supervised training scheme, and Stage 2 shows the unsupervised training scheme.

assumed to be closer to the real object shapes than BBs, are then post processed with Cues 1, which are adapted from Khoreva et al. (2017) to better suit histological data. The resulting segmentations are in turn used to generate training labels for the successive iteration.

**Cues 1** are namely the following:

- Any object missed completely during segmentation is reset to its corresponding BB. This avoids an increase of the instance of false negatives (FNs) with successive iterations.
- Similar to this, if the segmented object covers less than 50% of the original BB, it is also reset to the latter to avoid training with FN.
- If any pixel outside the BB of an object is marked as foreground, it is reset as background label because the BBs are assumed to be exhaustive. Thereby, we avoid training with a significant amount of false positives (FPs).

## 2.2. Stage 2 – Unsupervised learning

We extend the application of the approach further by utilizing the network trained in Stage 1 to facilitate completely unsupervised adaptation in the second step of our pipeline. Effectively, we obtain the training labels for Stage 2 by segmenting a new set of images, IS-train2 (see Fig. 2) using the network output by Stage 1. To keep the training scenario fair, IS-train2 is completely unseen in Stage 1 and we do not consider any GT annotations of IS-train2 data set during training.

For training we again evaluate a similar iterative approach as in Stage 1. However, here we cannot incorporate Cues 1 for the intermediate post processing because of the lack of GT. So, we rely on some basic statistical parameters (area and eccentricity) calculated from the training images

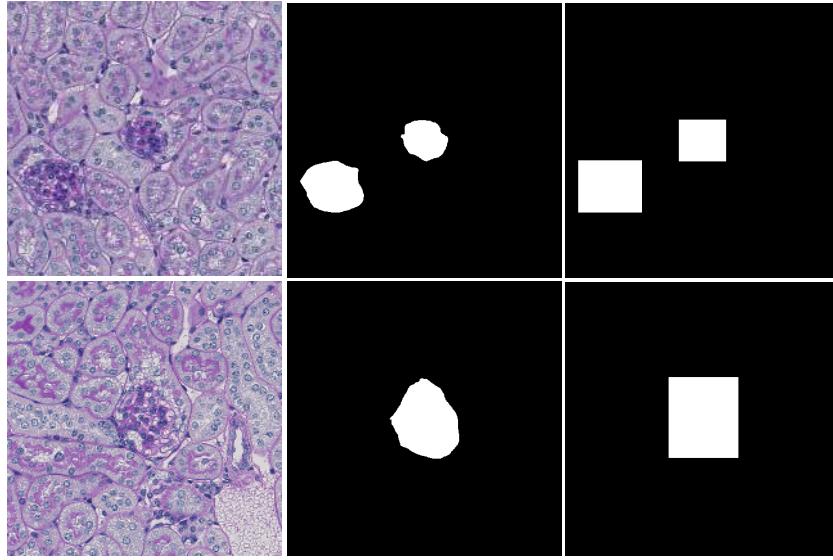


Figure 3: Example showing two  $500 \times 500$  size patches (left column), the corresponding GT masks (middle column), and the bounding boxes (BBs) (right column).

available for weakly supervised training, IS-train1 (see Fig. 2). For this purpose, we segment IS-train1 with the model resulting in Stage 1 because so far we only had BBs as the GT labels for these images. After segmentation, we obtain 'better than BB' segmentations' for the glomeruli on IS-train1, which are then used to calculate the distribution of their area and eccentricity. This information serves as Cues 2.

As explained in Stage 1, here we also train a model followed by segmentation and modification with the constraints (here, Cues 2) as one complete training round. As the initial labels in this stage were not GT, the statistics in Cues 2 was not expected to be a perfect representative of the typical size and shape of glomeruli. In other words, the statistical distribution obtained from these samples also included objects which can be considered as outliers regarding size or shape. Hence, it could be assumed to fit the distribution of the GT only roughly. Taking this into consideration, and to ascertain that unsupervised training incorporates as much correct information about the ROIs as possible, we ignored the objects lying in the marginal distributions (95% confidence interval). Hence, in this stage, after applying Cues 2, we retain only those objects in the segmented images which fall within the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentile of the respective distributions. This ensures that we retain only objects which are roughly the size and shape of glomeruli, and use these objects with a high confidence for further training.

### 3. Image dataset & experimental settings

The dataset used in this work consists of 22 WSIs of resected healthy mouse kidneys which are highly similar to human kidneys. The dataset was divided randomly into two parts, with five WSIs serving as test images (IS-test), and 17 as training images (eight for weakly supervised (IS-train1) and nine for unsupervised (IS-train2)). Image acquisition was performed with a Hamamatsu whole

slide scanner (NanoZommer 2.0-HT (C9600-13)) with a  $40 \times$  magnification. Each WSI was dyed with Periodic Acid-Schiff (PAS) stain and was roughly  $37,000 \times 37,000$  pixels in size. All the images are processed in the RGB color space.

To determine the optimal parameters for training in Stage 1, we performed some experiments with correctly labeled data for variable training set sizes. We considered  $2^3$  to  $2^8$  (8, 16, ..., 256) number of patches for training, extracted equally from a range of  $2^0$  to  $2^3$  (0, 2, ..., 8) number of WSIs. In all, we experimented with 24 different settings, beginning with the extraction of 8 to 256 patches from one WSI to extracting the similar sets of patches from eight WSIs. Here, we optimize the number of epochs based on the number of patches, with higher epochs for lesser training data. We increase the epochs from  $2^2$  to  $2^7$  with decreasing number of patches used for training. For each setting, we train the network for 10 iterations. We will refer to this as **Stage 0** in the remainder of the paper.

For **Stage 1**, we extract 16 patches each from eight WSIs (total 128 patches) for training with 32 epochs, similar to the previous experiment with the same number of patches. We train the network for 10 iterations. We then use the networks so obtained at their worst and best scores (networks N1 and N2, respectively) for performing experiments for **Stage 2**. That means we train two unsupervised networks as explained in Section 2.2 to evaluate the effect of Stage 1 on the performance of this step. We take five iterations into consideration for this stage.

## 4. Results

Fig. 5 shows the results obtained in all the stages based on five test images IS-test.

### Stage 0:

Fig. 5a shows the F-scores with their respective standard deviations for the experiments described in Stage 0. The individual curves for each WSI compare the F-scores reached with different number of patches. At the same time, the overlay of these curves allows the comparison of the effect of training with similar number of patches extracted from different number of WSIs.

Here we see that when training with eight WSIs, the F-score reaches a plateau ( $=0.89$ ) at 64 patches. Also noteworthy is the low standard deviations in all cases of #WSI=8. On the other hand, when using only one WSI, low F-scores are exhibited with high standard deviations.

When increasing the training data set size in terms of the number of patches extracted from each WSI, the F-scores do not always show an increment. Nevertheless, in except three of the 24 tested settings, we get a mean F-score of above 0.75 even when training the network with as low as eight annotations. However, the standard deviations in all cases are much higher compared to the setting of eight WSIs.

### Stage 1:

Fig. 4 shows the results of segmentation after every iterative training of the network in this stage. The figure shows an overlay of the segmentation masks and the GT in one WSI. A perfect overlay of the two is shown in green, falsely segmented objects are red, while (partly) missed objects are shown in blue. A quick glance shows that the FPs reduce greatly with more iterations, as the number of instances of FNs increases.

Fig. 5b reports the mean F-scores, precision and recall along with their standard deviations for Stage 1. The graph illustrates the values reached on segmenting the test images with the networks obtained after each iteration (total 10) of the training procedure.

Here, we notice that although the recall drops consistently until the sixth iteration, the precision increases, leading to a rise in the F-score. The F-score across all the iterations reaches a peak at this point in the plot.

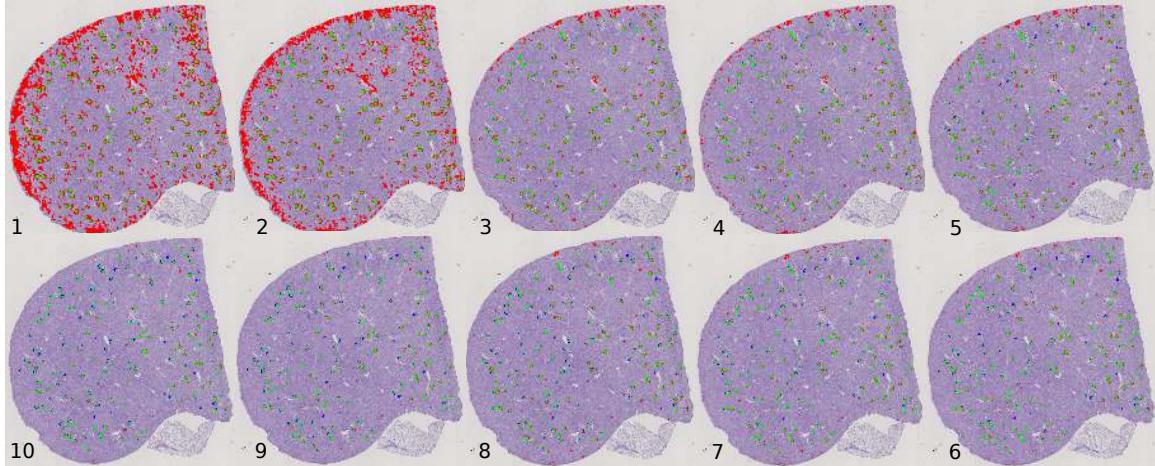


Figure 4: Example showing how the segmentation results evolve on one WSI from 1 to 10 iterations in Stage 1 (clockwise from the top left). Annotation legend: red=FP, blue=FN, green=TP.

### Stage 2:

In Fig. 5c and Fig. 5d, we see the F-score, precision and recall values for the experiments of Stage 2. Here, we see the results for the Stage 2 network (iteration 1 to 5) trained based on the output from the network from the first (N1) and sixth (N2) iteration of Stage 1 (iteration 0), respectively.

If training Stage 2 after running the first and sixth iterations of Stage 1, respectively, we notice a strong improvement in the F-score between the 0<sup>th</sup> and 1<sup>st</sup> iteration. Iteration one here is effectively the the first iteration of Stage 2. Although not much improvement in the values is achieved with more iterations, it is important to note that the standard deviations decrease slightly.

## 5. Discussion

Referring to Fig. 5a, we may comment that the highest F-scores are obtained with eight WSIs because in this setting we incorporate inter-slide variability much more effectively than in the other settings with fewer number of WSIs. This is reflected also by the comparatively lower standard deviations in the former case.

With these observations, we safely chose WSI=8, patches=16 (total patches=128) as the most optimal configuration for Stage 1 (F-score=0.89).

The scores of Stage 1 (Fig. 5b) may be accounted for as follows. As we initiate training in this stage with BBs, the first iteration has comparatively high number of FPs. This explains the

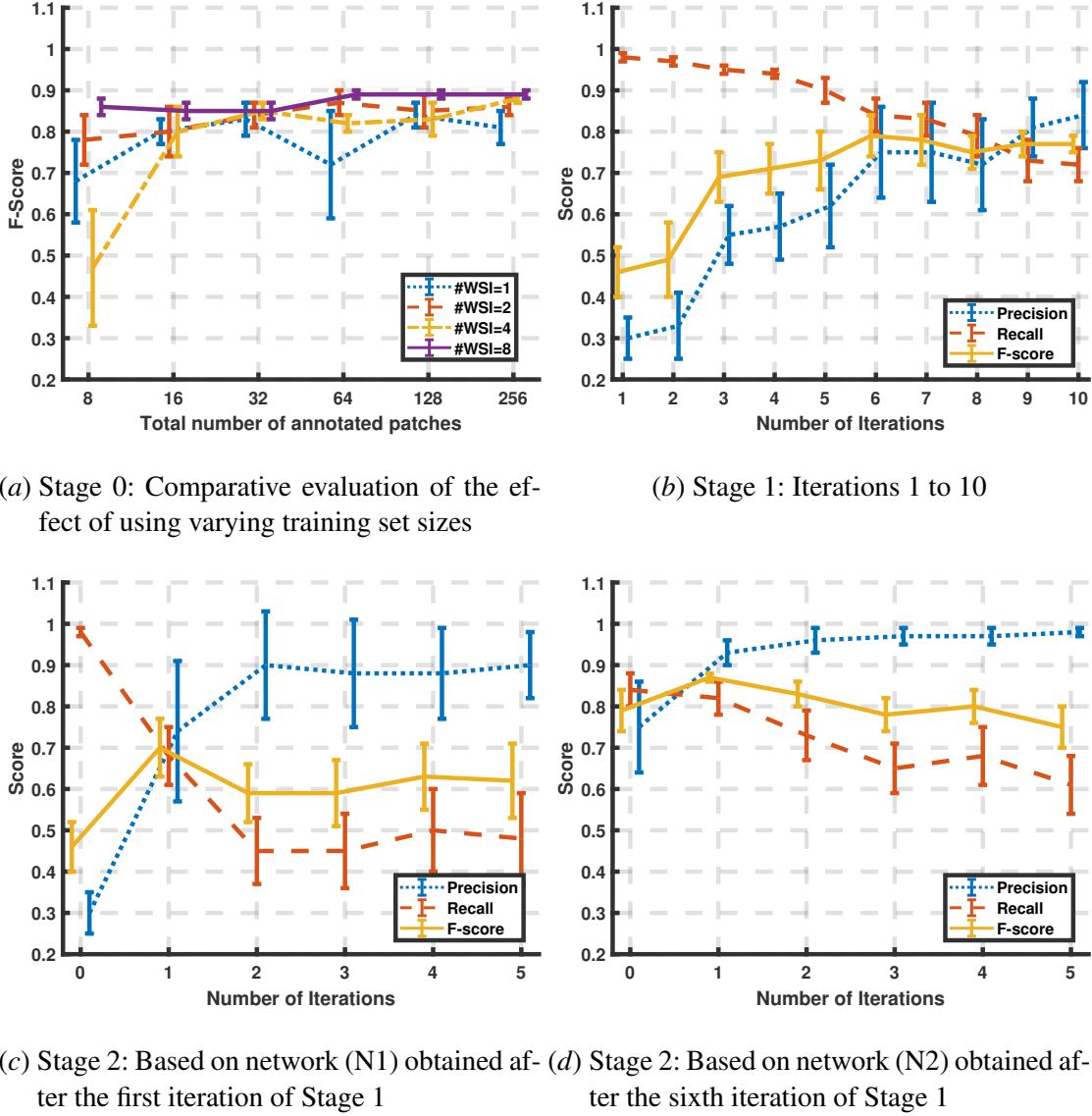


Figure 5: Segmentation results on five test images (IS-test) for Stages 0, 1 and 2.

low precision, yet high recall because the network was essentially trained with all the TPs. With each iteration, as the annotations get closer to the GT, precision keeps improving while recall is decreasing. We observe that the overall F-scores increase steadily and the highest score (0.79) is reached at the sixth iteration, when the precision and recall both fall gradually. This is because the iterative training is then gradually incorporating more and more FNs.

The results of Stage 2 (Fig. 5c and 5d) show that the network benefits from training on the fully-automatically generated training data because the F-scores improve strongly in the first iteration. However, with the succeeding iterations, the performance decreases. A reason for this could be that Cues 2 used as constraints in this stage were not suitable for the task.

As a qualitative comparison between the annotation efforts with the new approach and the supervised training method, we compare the time taken for labeling in both the cases. For this purpose, we evaluate the experiment setting requiring 16 annotated patches (considering one glomerulus per patch) each, on eight WSIs. The average time taken to label a glomerulus precisely was noted to be about 60 seconds, and to draw a BB around it, approximately 10 seconds. For the proposed method (BB annotations) we needed only 21 minutes (approximately), while for the supervised method (precise annotations), we needed 128 minutes ( $> 2$  hours). Our method saves annotation time significantly (6.5 times), which is an important consideration especially for medical data.

## 6. Conclusion

Automation of histopathological image analysis procedures typically demands a lot of manually annotated data, which is difficult and time-consuming to obtain. In this work we seek to minimize the limitations caused by the unavailability of fully labeled data by adopting a weakly supervised approach, whereby we require only limited and coarse annotations. This effectively means that we work with imprecise easy-to-collect labels (BBs), refining them with prior knowledge about the dataset, which is easily available. In this endeavor, we achieve mean F-scores of 0.79 when training with as low as eight sparsely and imprecisely annotated WSIs. We address another major limitation in histological image analysis. Although substantial amount of image data is often available, it is not routinely utilized. We exploit such unlabeled data to further improve the performance of our weakly supervised models and achieve accuracy values (F-score=0.87) comparable to fully supervised models (F-score=0.89). It is also noteworthy that these results are achieved with significantly reduced annotation effort and time.

## Acknowledgement

This work was supported by the German Research Foundation (DFG) under grant no. ME3737/3-1.

## References

- Michael Gadermayr, Vitus Appel, Barbara M. Klinkhammer, Peter Boor, and Dorit Merhof. Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'18)*, 2018.
- Michael Gadermayr, Ann-Kathrin Dombrowski, Barbara Mara Klinkhammer, Peter Boor, and Dorit Merhof. Cnn cascades for segmenting sparse objects in gigapixel whole slide images. *Computerized Medical Imaging and Graphics*, 2019.
- N. Herve, A. Servais, E. Thervet, J.-C. Olivo-Marin, and V. Meas-Yedid. Statistical color texture descriptors for histological images analysis. In *Proceedings of ISBI'11*, pages 724–727, 2011.
- Tsuyoshi Kato, Raissa Relator, Hayliang Ngouv, Yoshihiro Hirohashi, Osamu Takaki, Tetsuhiro Kakimoto, and Kinya Okada. Segmental HOG: new descriptor for glomerulus detection in kidney microscopy image. *BMC Bioinformatics*, 16(1), 2015.

Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017.

Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Aided Interventions (MICCAI'15)*, pages 234–241. Springer International Publishing, 2015.

# Generative Image Translation for Data Augmentation of Bone Lesion Pathology

**Anant Gupta\***

*Courant Institute, New York University*

ANANTGUPTA@NYU.EDU

**Srivas Venkatesh**

**Sumit Chopra**

**Christian Ledig**

*Imagen Technologies, Inc., New York*

SRIVAS@IMAGEN.AI

SUMIT@IMAGEN.AI

CHRISTIAN@IMAGEN.AI

## Abstract

Insufficient training data and severe class imbalance are often limiting factors when developing machine learning models for the classification of rare diseases. In this work, we address the problem of classifying bone lesions from X-ray images by increasing the small number of positive samples in the training set. We propose a generative data augmentation approach based on a cycle-consistent generative adversarial network that synthesizes bone lesions on images without pathology. We pose the generative task as an image-patch translation problem that we optimize specifically for distinct bones (humerus, tibia, femur). In experimental results, we confirm that the described method mitigates the class imbalance problem in the binary classification task of bone lesion detection. We show that the augmented training sets enable the training of superior classifiers achieving better performance on a held-out test set. Additionally, we demonstrate the feasibility of transfer learning and apply a generative model that was trained on one body part to another.

**Keywords:** Bone lesion, X-ray, generative models, data augmentation.

## 1. Introduction

Deep neural networks have demonstrated their potential to reach human-level performance for image classification, however, their performance generally correlates with the amount of available samples (Domingos, 2012). When focusing on rare medical conditions, the limited availability of pathological (positive) training images can cause severe class imbalance that limits the accuracy of these models. In contrast, the collection of normal (negative) cases is often substantially simpler. One example of a pathology that is both of high interest but also rare are bone lesions (Franchi, 2012). The classification of the presence of bone lesion pathology in X-ray images is the subject of our work.

Traditional methods to handle class-imbalance, such as image transformations (Hussain et al., 2017) and different sampling strategies (Li et al., 2010; Dubey et al., 2014), are often of limited benefit as they do not address the inherent problem of dealing with a small training set not fully representing the underlying data distribution. Recent works have proposed the use of synthetic data in order to augment and increase diversity in the training set (Antoniou et al., 2017; Mariani et al., 2018). However, learning to generate high-resolution images from random noise requires an often prohibitively large training dataset.

---

\* Work done when author was at Imagen Technologies, Inc.

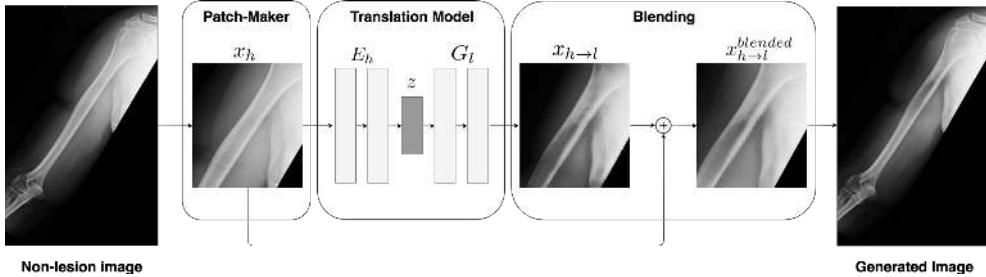


Figure 1: Pipeline of the lesion generation process on non-lesion images.  $x_h$  is non-lesion patch;  $E_h$ ,  $z$  and  $G_l$  are non-lesion encoder, latent representation and lesion generator respectively.  $x_{h \rightarrow l}$  is the generated lesion and  $x_{h \rightarrow l}^{blended}$  is the result after alpha-blending.

In this work, we aim to synthesize bone lesions by translating spatially-constrained patches extracted from non-pathological X-ray images rather than generating from scratch. The translation task can be defined as modifying an image patch that was sampled from a source data distribution to a target data distribution. In this work, it corresponds to sampling a patch from healthy(non-lesion) images and translating it to a patch that is characteristic of bone lesion images. The lesion-generation pipeline is illustrated in Figure 1. The model is trained on patches to ensure localized generation of pathology. A blending approach merges the translated patches back into those full-images. A subset of the generated images is filtered to form the augmented training set by performing pseudo-labelling. We observed non-trivial performance gains in the task of bone lesion detection for individual body parts (humerus, tibia, femur) when trained using this augmented set. We further show that transfer learning can be a viable option to enhance the training set of body parts for which a powerful image-translation model cannot be trained due to insufficient or noisy samples.

## 2. Related Work

Data augmentation is a well-studied problem in machine learning. Employing transformation-based augmentation techniques (Rajkomar et al., 2017; Kohli et al., 2017) or transfer learning by using pretrained weights, are common approaches (Rajkomar et al., 2017), which are used in this work as well.

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have been used successfully in the medical imaging domain to accomplish tasks such as image translation (Wolterink et al., 2017; Nie et al., 2017), segmentation (Xue et al., 2018; Kamnitsas et al., 2017) and data augmentation. Shin et al. (2018) generate brain tumors for data augmentation by translating segmentation masks to multi-parameteric magnetic resonance (MR) images, using a multi-modal dataset with uniform view. Frid-Adar et al. (2018) use a small dataset of regions of interest of liver lesions in CT images to train a DCGAN (Radford et al., 2015) and generate an augmented set. In comparison, our method focuses on generative data augmentation using a small number of high-resolution X-ray images often varying in positional view even within a single body part.

Salehinejad et al. (2018) use DCGAN to generate chest X-rays for multiple pathologies. Plausible samples are filtered out by a team of radiologists to create an augmented set. In our work we

perform filtering in an automated manner to mine for hard positives. Recent work by [Lau et al. \(2018\)](#) generate scars on cardiac MR scans and employ a blending mask to remove unwanted artifacts. To the best of our knowledge, there is no existing literature that addresses the problem of bone lesion classification by automatically generating pathology in normal radiographs to enhance the training set.

### 3. Methodology

#### 3.1. Unsupervised Image-to-Image Translation Model

The generation of bone lesions is posed as an unsupervised image to image translation task ([Liu et al., 2017](#)). In this task,  $P\chi_l(x_l)$  and  $P\chi_h(x_h)$  are two marginal distributions from which X-ray patches of bones with lesion  $x_l$ , and non-lesion  $x_h$  are drawn respectively. The model maps these samples to a shared latent representation, using encoders for respective distributions:  $E_l(x_l) = E_h(x_h) = z \in \mathcal{Z}$ . The generators respective to each distribution decode back the input sample from this latent vector:  $G_l(z) = x_l, G_h(z) = x_h$ . Lesion-like properties are generated in a normal bone X-ray with the following translation operation:  $G_l(E_h(x_h)) = x_{h \rightarrow l}$ . This framework is based on the assumption that there exists an unknown but finite joint distribution  $P_{\chi_l, \chi_h}$ , which the shared latent space can learn to represent.

In order to find the optimal hypothesis for this problem, the lesion encoder-generator is a variational autoencoder (VAE) ([Kingma and Welling, 2013](#)), whose loss function maximizes the Evidence Lower Bound (ELBO) by minimizing the following objective:

$$\mathcal{L}_{VAE_l} = \lambda_1 \text{KL}(q_l(z_l|x_l)||\mathcal{N}(z|0,I)) - \lambda_2 \mathbb{E}_{z_l \sim q_l(z_l|x_l)} [\log p_{G_l}(x_l|z_l)] \quad (1)$$

where  $q_l(z_l|x_l)$  is the distribution from which  $z_l$  (encoding of  $x_l$ ) is sampled. In the first term, the KL divergence between this distribution and the prior is minimized, which encourages  $q_l(z_l|x_l)$  to follow a normal (zero-mean, unit-covariance) distribution. The second term aims to maximize the log-likelihood of  $p_{G_l}$ . The same formulation is followed to train a second VAE for normal, non-lesion samples. This would ensure each generator is able to reconstruct images of the respective distribution.

An adversarial objective ([Goodfellow et al., 2014](#)) is employed to help in learning to translate from one domain to another. In this setting, the lesion generator  $G_l$  is conditioned on the latent encoding of a healthy patch  $z_h$  and the generated sample is evaluated by the discriminator  $D_l$  to classify whether the sample was drawn from  $P\chi_l(x_l)$  or not. This encourages the generator to create lesion-like image features while constructing an image sample. The GAN objective is defined as:

$$\mathcal{L}_{GAN_l} = \lambda_0 [\mathbb{E}_{x_l \sim P_{\chi_l}} [\log D_l(x_l)] + \mathbb{E}_{z_h \sim q_h(z_h|x_h)} [\log(1 - D_l(G_l(z_h)))] ] \quad (2)$$

The conceptual shared latent space is implemented in practice by weight-sharing across the two VAEs. The shared latent space implies a cycle-consistency constraint ([Zhu et al., 2017](#)) that ensures successful circular back and forth mapping between domains:

$$\begin{aligned} \mathcal{L}_{CC_l} = & \lambda_3 [\text{KL}(q_l(z_l|x_l)||\mathcal{N}(z|0,I)) + \text{KL}(q_h(z_h|x_{l \rightarrow h})||\mathcal{N}(z|0,I))] \\ & - \lambda_4 \mathbb{E}_{z_h \sim q_h(z_h|x_{l \rightarrow h})} [\log p_{G_l}(x_l|z_h)] \end{aligned} \quad (3)$$

This objective aims at preserving the original information of the input image and prevents mode collapse by translating all images to a single output image. Similar loss objectives are minimized

for  $\text{VAE}_h$ ,  $\text{GAN}_h$  and  $\text{CC}_h$ . The hyperparameters ( $\lambda$ ) control the contribution of each respective loss function. The network is jointly trained to optimize the following objective:

$$\min_{E_l, E_h, G_l, G_h} \max_{D_l, D_h} \mathcal{L}_{\text{VAE}_l} + \mathcal{L}_{\text{GAN}_l} + \mathcal{L}_{\text{CC}_l} + \mathcal{L}_{\text{VAE}_h} + \mathcal{L}_{\text{GAN}_h} + \mathcal{L}_{\text{CC}_h} \quad (4)$$

### 3.2. Patch-making

Bone lesions tend to cause local alterations in bone anatomy without substantially affecting the global visual appearance of the image. We therefore aim to translate localized image patches rather than training a translation model for the complete images. This technique has the following advantages: i) computationally cheaper, ii) multiple patches can be created from a single image, iii) lesion-like features are more prominent on the localized patch, which supports efficient training of the translation model. Lesion patches are created by randomly cropping a square area (if image size permits) by a factor  $s \in \{1, 2\}$  larger than the larger side of the bounding box around the area containing the pathology. This area is marked with a manually annotated bounding box (c.f. Figure 2). We employ an heuristic to automate cropping of normal patches. We identify potential ‘crop-areas’ in a two step process. First we randomly choose  $n$  similar non-lesion images for each lesion image. Second we crop each non-lesion image based on the lesion annotations of the matched lesion image. All non-lesion patches with a mean, normalized ([0,1]) pixel intensity of less than 0.15 are assumed to not contain bone structure and are dropped from the dataset.

### 3.3. Blending

The translated patches also exhibit subtle changes in the overall image characteristics, such as contrast and brightness. This leads to the patch being visibly distinct when placed back in the full-image after translation. We employ alpha-blending to smoothly blend the translated patch in the original image:  $x_{h \rightarrow l}^{\text{blended}} = \alpha x_{h \rightarrow l} + (1 - \alpha)x_h$ . Specifically, we define a locally varying blending factor  $\alpha$  as:  $\alpha = \cos(|i|^n * \frac{\pi}{2}) \cos(|j|^n * \frac{\pi}{2})$ , where  $i$  and  $j$  are evenly spaced linear interpolations between [-1,1] along each dimension of the 256x256 square mask and  $n$  is a hyper-parameter, with  $n = 1$  being chosen based on the qualitative analysis of the resulting blended image.

### 3.4. Pseudo-Labeling

We aim to augment the training set with images containing a prominent lesion after the blending operation. We perform hard positive mining (Lee, 2013) on the generated set using a classifier trained on the available empirical training data (baseline). Based on a threshold parameter  $t$ , the baseline classifier segregates the generated samples into two disjoint sets: samples with extreme lesion-like properties, and noisy samples. The former is used for augmentation and added to the training set.

## 4. Experimental Setup

### 4.1. Dataset

A set of adult X-ray images showing bone anatomy with and without lesion are sourced from various U.S. hospitals and assessed by expert, board-certified radiologists by drawing bounding boxes around the target pathology (bone lesions) of concern (c.f. Figure 2). A test dataset is held out

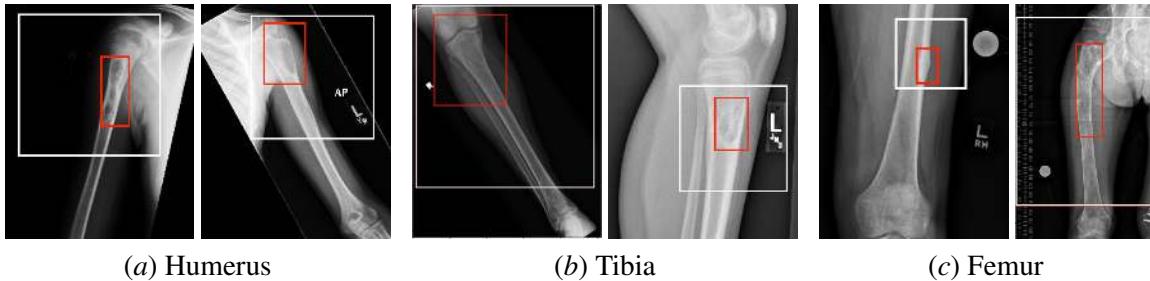


Figure 2: Bone lesion with expert-annotated bounding box (red). Random sized patches (white), cropped around lesions, used for training the generative model.

containing sufficient positive samples for evaluation and used at no point to train or fine-tune any model. The remaining dataset is then used for training and validating both the classifiers and the translation models. Full-images are used for the classification task whereas the translation task is performed on image patches.

**Classification Task:** Images with presence of confounding features (e.g., congenital deformity, fixation hardware) negatively impacted the model’s classification performance. We thus removed those images from all datasets when training classification models. A summary of the data split, excluding augmented samples, is provided for the three investigated body parts in Table 1 (left). The generated images used for augmentation are only added to the training set but not the validation set.

**Translation Task:** We do not remove images from the lesion set when training the generative model, as it is trained on cropped image patches that are less affected by the confounding features. However, we remove images with confounding features from the non-lesion set to ensure that the augmented training set does not contain confounding images. The class split is kept balanced to facilitate training of the models. The images from the negative class in the training set that are not used to train the generative model are used for creating the augmented training set. Image patches are cropped from those images as described in Section 3.2. Table 1 (right) reports the distribution of the patch dataset and the configuration settings.

#### 4.2. Model Architecture and Optimization

The classifier is a dilated residual net (DRN) (Yu et al., 2017). Dilated convolutional filters increase the receptive field of view and help capture finer details in high-resolution images. Images are downsampled to 1024x512 pixels in our experiments. To avoid overfitting on our comparatively small training set, our model was pretrained on a larger corpus of X-ray images for the auxiliary task of fracture detection. Training the classifier in this work involves fine-tuning of the last two convolutional blocks and the fully-connected layer of the model. Regularization is performed through augmentation procedures including linear transformations, along with weight decay. The model is optimized using Adam with an initial learning rate of 0.0001 which decays by a factor of 0.9 when the performance on the validation set plateaus.

The variability of the body part specific bone anatomy influenced our ability to train the translation model. Models on more diverse datasets like tibia could only be trained if the patch sizes were not larger than the bounding box ( $s = 1$ ). On the other hand, a comparably uniform anatomical

Table 1: Datasets for each model (ratio denotes lesion:non-lesion class split). Left: images used for classification. Right: Extracted patches used for generation. Source samples are only non-lesion and used for creating the augmented sets.  $s$  is the factor by which the patch is larger than the larger side of the bounding box.  $n$  is the number of non-lesion images chosen against each lesion image.

Classification Task				Translation Task				
Body part	Train	Val	Test	Body part	Train	Source	$s$	$n$
Humerus	268:2295	41:305	50:500	Humerus	536:536	4643	2	10
Tibia	214:14482	22:1628	50:500	Tibia	515:515	4680	1	7
Femur	32:4558	14:573	50:500	Femur	285:285	9171	2	10



Figure 3: Stages of patch translation for full-image (top) and selected patch, highlighted with white box (bottom): i) original, ii) translated and iii) blended.

view among humerus images allowed training with larger patch sizes ( $s = 2$ ). The adversarial loss weight influenced the qualitative results. Setting  $\lambda_0 = 1$  resulted in a change in texture of the bone, rather than synthesis of a circular lesion. The model architecture and loss weighting which yielded the best results is directly adapted from (Liu et al., 2017) without any major modifications. We found residual connections in the encoder and generator beneficial and hypothesize that copying the common features in the patch helps in training on such a small dataset. Figure 3 demonstrates the blending process after translation using the default mask.

### 4.3. Transfer Learning

In comparison to the available humerus X-rays, the available tibia and femur datasets were highly heterogeneous. We observed highly variable radiographic views and frequent confounding image content (e.g. external objects) in the not excluded positives. This made it particularly challenging to train a valuable generative model for tibia and unfeasible for femur, regardless of the patch size.



Figure 4: Bone lesion generated using transfer learning techniques. The variation in positional views within each body part makes it challenging to train a generative model.

Table 2: Ablation study of  $t$  (threshold score of pseudo-labeller) reporting classifier performance on humerus test-set. Sensitivity and Specificity are calculated at the OP. Significantly different AUC with respect to baseline indicated with \*.

Type	$t$	Augmented Samples	ROC AUC (CI 95%)	Sens.	Spec.	OP
Baseline	0	0	0.876 (0.817-0.926)	0.9	0.776	0.455
Augmented	0.70	1412	0.882 (0.829-0.928)	0.80	0.842	0.390
	0.85	577	0.899 (0.854-0.939)	0.82	0.802	0.086
	0.90	401	<b>0.924 (0.889-0.955)*</b>	0.84	0.798	0.058
	0.95	257	0.877 (0.820-0.926)	0.90	0.766	0.273

Transfer learning, being a well known method to make use of learned representations from one domain and leveraging those in another, was seen as a suited technique to circumvent the issues around training tibia and femur models. We explored the potential of using transfer learning by i) employing the translation model trained on humerus to generate lesions on other body parts, ii) doing pseudo-labelling based on the humerus baseline classifier. For tibia we set  $s = 1$  to keep it consistent with the tibia-specific generative model. For femur we set  $s = 2$  to keep it consistent with the humerus configuration.

#### 4.4. Performance Measures

We report the Area Under the ROC-Curve (AUC) and the bootstrapped 95% Confidence Interval (CI). It was ensured that all models are compared on the same set of bootstrap samples. This allows us to examine the bootstrap-wise difference in AUC scores of models against the baseline. We consider a model to be significantly different to the baseline if the 95% CI of those bootstrapped difference scores does not contain zero. We report Sensitivity (Sens) and Specificity (Spec) by defining an Operating Point (OP) over the validation set as the point which minimizes  $(1 - \text{true positive rate})^2 + (\text{false positive rate})^2$  over the ROC curve. We focus on AUC scores since the operating point is, due to the low sample size of our validation set, highly variable and does not generalize well across experiments.

Table 3: Comparison of classifier model performance on tibia and femur test-sets. A translation model couldn't be trained for femur due to high diversity of radiographic view and insufficient samples.  $TL_G$ =Inference with humerus translation model,  $TL_{PL}$ =Pseudo-labelling with humerus baseline model.

(a) Tibia						
Augment Type	Augmented Samples	ROC AUC (CI 95%)	Sens.	Spec.	OP	
Baseline	0	0.618 (0.532-0.705)	0.54	0.652	0.300	
Augmented	124	0.640 (0.547-0.732)	0.6	0.542	0.244	
$TL_G$	118	0.642 (0.550-0.735)	0.52	0.66	0.290	
$TL_G + TL_{PL}$	1264	<b>0.698 (0.610-0.785)*</b>	0.74	0.464	0.066	

(b) Femur						
Augment Type	Augmented Samples	ROC AUC (CI 95%)	Sens.	Spec.	OP	
Baseline	0	0.533 (0.441-0.627)	0.64	0.376	0.010	
$TL_G$	579	0.601 (0.504-0.695)*	0.56	0.61	0.012	
$TL_G + TL_{PL}$	1342	<b>0.682 (0.594-0.764)*</b>	0.66	0.67	0.008	

## 5. Results

The augmentation set is composed of generated images that the baseline classifier assigns a confidence score of  $t$  or higher. In the transfer learning setting, the humerus baseline classifier is used to select generated images for tibia and femur respectively. A grid search is performed on the validation set and  $t$  is chosen to be the value that gives the highest validation set AUC ( $t_{\text{humerus}} = 0.9$ ,  $t_{\text{tibia}} = 0.9$ ,  $t_{\text{femur}} = 0.95$ ). To assess the influence of this parameter we report AUCs on the humerus test set for different values of  $t$  in Table 2. We observe that the approach is sensitive to the choice of  $t$  which, however, can be successfully chosen on the validation set. Adding either insufficient number of samples (larger  $t$ ) or excessive low-quality samples (smaller  $t$ ) reduces the benefit of data augmentation. We observed a significant increase in AUC of around 5% over the humerus baseline model at  $t = 0.9$ , as determined on the validation set.

For tibia we observed similar minor improvements ( $\approx 2\%$ ) when using either the humerus or tibia generative model. However, when further relying on the humerus baseline classifier for sample selection we observed a more substantial performance gain of around 8% that was borderline to significant in the conducted test. For femur we observed significant gains in AUC when employing transferring knowledge from the humerus models. In particular, we observed an substantial improvement of around 15% over the barely discriminative femur baseline classifier. See Table 3 for the full quantitative analysis for tibia and femur when using transfer learning. Figure 4 illustrates some of the generated samples for tibia and femur obtained using transfer-learning based on the humerus model.

## 6. Conclusion

We trained a generative model that can represent some properties of the target pathology (bone lesions in X-ray) and synthesize those into sample patches drawn from another distribution (normal anatomy). When employing generative models for augmenting medical datasets, great care needs to be taken to avoid and control for possibly introduced bias. Future work should be concerned with the exploration of those limitations and explore the method’s potential on both a more diverse set of disease pathology and other modalities.

## Acknowledgements

The project is funded by Imagen Technologies, Inc. The work presented in this manuscript is for research purposes only and is not for sale within the United States.

## References

- Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.
- Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.
- Rashmi Dubey, Jiayu Zhou, Yalin Wang, Paul M Thompson, Jieping Ye, Alzheimer’s Disease Neuroimaging Initiative, et al. Analysis of sampling techniques for imbalanced data: An n= 648 adni study. *NeuroImage*, 87:220–241, 2014.
- Alessandro Franchi. Epidemiology and classification of bone tumors. *Clinical Cases in mineral and bone metabolism*, 9(2):92, 2012.
- Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *arXiv preprint arXiv:1803.01229*, 2018.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- Zeshan Hussain, Francisco Gimenez, Darvin Yi, and Daniel Rubin. Differential data augmentation techniques for medical imaging classification tasks. In *AMIA Annual Symposium Proceedings*, volume 2017, page 979. American Medical Informatics Association, 2017.
- Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging*, pages 597–609. Springer, 2017.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

- Marc D Kohli, Ronald M Summers, and J Raymond Geis. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 c-mimi meeting dataset session. *Journal of digital imaging*, 30(4):392–399, 2017.
- Felix Lau, Tom Hendriks, Jesse Lieman-Sifry, Sean Sall, and Dan Golden. Scargan: Chained generative adversarial networks to simulate pathological tissue on cardiovascular mr scans. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 343–350. Springer, 2018.
- Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013.
- Der-Chiang Li, Chiao-Wen Liu, and Susan C Hu. A learning method for the class imbalance problem with medical data sets. *Computers in biology and medicine*, 40(5):509–518, 2010.
- Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.
- Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018.
- Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 417–425. Springer, 2017.
- Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- Alvin Rajkomar, Sneha Lingam, Andrew G Taylor, Michael Blum, and John Mongan. High-throughput classification of radiographs using deep convolutional neural networks. *Journal of digital imaging*, 30(1):95–101, 2017.
- Hojjat Salehinejad, Shahrokh Valaee, Tim Dowdell, Errol Colak, and Joseph Barfett. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 990–994. IEEE, 2018.
- Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Sjenem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 1–11. Springer, 2018.
- Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Išgum. Deep mr to ct synthesis using unpaired data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 14–23. Springer, 2017.
- Yuan Xue, Tao Xu, Han Zhang, L Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scale l1 loss for medical image segmentation. *Neuroinformatics*, pages 1–10, 2018.

Fisher Yu, Vladlen Koltun, and Thomas A Funkhouser. Dilated residual networks. In *CVPR*, volume 2, page 3, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.

# Cluster Analysis in Latent Space: Identifying Personalized Aortic Valve Prosthesis Shapes using Deep Representations

**Jannis Hagenah<sup>1</sup>**

HAGENAH@ROB.UNI-LUEBECK.DE

**Kenneth Kühl<sup>1</sup>**

<sup>1</sup> Institute for Robotics and Cognitive Systems, University of Lübeck, Retzeburger Allee 160, 23562 Lübeck, Germany

**Michael Scharfschwerdt<sup>2</sup>**

MICHAEL.SCHARFSCHWERDT@UKSH.DE

<sup>2</sup> Department of Cardiac Surgery, University Hospital Schleswig-Holstein, Ratzeburger Allee 160, 23562 Lübeck, Germany

**Floris Ernst<sup>1</sup>**

ERNST@ROB.UNI-LUEBECK.DE

## Abstract

Due to the high inter-patient variability of anatomies, the field of personalized prosthetics gained attention during the last years. One potential application is the aortic valve. Even though its shape is highly patient-specific, state-of-the-art aortic valve prostheses are not capable of reproducing this individual geometry. An approach to reach an economically reasonable personalization would be the identification of typical valve shapes using clustering, such that each patient could be treated with the prosthesis of the type that matches his individual geometry best. However, a cluster analysis directly in image space is not sufficient due to the curse of dimensionality and the high sensitivity to small translations or rotations. In this work, we propose representation learning to perform the cluster analysis in the latent space, while the evaluation of the identified prosthesis shapes is performed in image space using generative modeling. To this end, we set up a data set of 58 porcine aortic valves and provide a proof-of-concept of our method using convolutional autoencoders. Furthermore, we evaluated the learned representation regarding its reconstruction accuracy, compactness and smoothness. To the best of our knowledge, this work presents the first approach to derive prosthesis shapes data-drivenly using clustering in latent space.

**Keywords:** personalized medicine, representation learning, aortic valve, personalized prosthetics, unsupervised learning

## 1. Introduction

The geometry of the aortic valve is highly patient-specific. Especially the three valve leaflets differ in their size and shape, while the interdependency of these shapes define the correct function of the valve (De Kerchove et al., 2017). However, state-of-the-art valve prostheses are not capable of remodeling this individual shape. While mechanical prostheses are most commonly designed with only two leaflets or without any leaflet at all, the three leaflets of biological prostheses have the same shape, resulting in a radial symmetry of the valve that is barely found in nature (Pibarot and Dumesnil, 2009). During the last years, multiple studies indicated the high impact of the aortic valve geometry on the whole circulatory system, including the brain circulation (Al-Atassi et al., 2015), (Blais et al., 2003). Hence, it can be assumed that a personalization of aortic valve prostheses could increase the patient’s outcome significantly. Furthermore, it could be shown that a valve that

matches the anatomy of the surrounding tissue has a lower risk of cavitations and reduces the stress inside the leaflet material ([Andersen et al., 2006](#)). These effects would lead to a longer lifetime of personalized prostheses and accordingly to a reduced risk of follow-up surgeries.

Even though imaging of the leaflets is a remaining challenge due to the thin structure and the high movement, it could be shown that an estimation of the individual planar leaflet shapes is possible just based on an ultrasound image of the aortic root, utilizing Support Vector Regression ([Hagenah et al., 2018a](#)). Hence, a personalization of the leaflet shapes in a prosthesis is possible in general. However, a completely individual manufaction of personalized prostheses is unrealistic in the near future due to economical, logistical and regulatory issues. An alternative approach for a trade-off between these issues and a higher patient outcome would be to offer a number of specific prostheses types that approximate the realistic distribution of valve shapes. Then, each patient could be treated with the prosthesis type that matches his or her individual anatomy and physiology best. This leads to a classification problem, which should be easier to solve than the regression problem from ([Hagenah et al., 2018a](#)) and presents a cost-efficient and hence realistic way of aortic valve prostheses personalization.

In this paper, we present a method to perform a cluster analysis in aortic valve leaflet shapes to identify these valve types. Unfortunately, a clustering in image space where each pixel presents one dimension is not suitable due to an effect called the curse of dimensionality ([Keogh and Mueen, 2011](#)). In high dimensional spaces, distance metrics are influenced by the high number of dimensions and variance in the data can not be incorporated in a sufficient way. Additionally, small translations or rotations of the valve might lead to big distances in image space, even though the shape of the valve stays the same. Hence, we propose to perform the cluster analysis in a latent space description of the valves using representation learning. Utilizing generative modeling, the latent cluster centers can be transformed back to image space. Like this, the cluster analysis focuses on abstract, highly descriptive features instead of pixel-wise grayscale values, while the evaluation of the identified prosthesis shapes can be done using intuitive metrics in image space. To this end, we set up a sufficient data set of pig heart valves and performed a proof-of-concept study using convolutional autoencoders. We analyzed the influence of the number of identified valve types on the capability of reproducing all individual valve shapes that are present in our data set. Furthermore, we performed an analysis of the networks hyperparameters and evaluated the identified valve representation regarding its compactness, accurateness and smoothness.

### **1.1. Contribution of this work**

The contribution of this work is two-fold. First, our resulting model presents the first data-driven typisation of aortic valve shapes regarding personalized medicine. Second, to the best of our knowledge, the proposed method is the first approach to identify prosthesis shapes using clustering in a latent space description in general. It is transferable to comparable problems and could be the basis for a new subfield of personalized prosthetics.

## **2. Material and Methods**

This section is divided into three parts. First, the generation of the porcine dataset and the preprocessing pipeline is described (adapted from ([Hagenah et al., 2018b](#))). Second, since a clustering directly in the image space is not feasible, an autoencoder is used to learn a compact representation

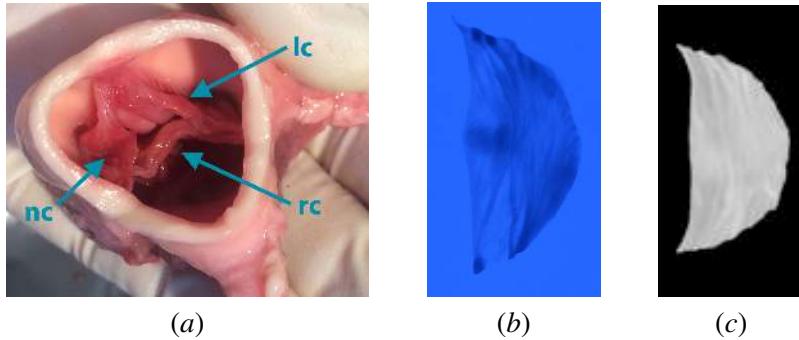


Figure 1: Data set generation. (a) Extracted aortic root from above with the right-coronary (rc), left-coronary (lc) and non-coronary (nc) leaflet. (b) Raw photo of one leaflet with blue backlight illumination. (c) Processed leaflet image.

of the valve geometry. In the third subsection, the clustering method as well as the evaluation of the identified prosthesis shapes in image space is presented.

## 2.1. Data Set Generation and Preprocessing

Due to their very thin structure and their fast movement during the cardiac cycle, detailed imaging of the leaflets is a tough task. Hence, we bypass this problem by cutting out the aortic valve leaflets of fresh porcine aortic roots, assuming that the shape of the aortic valve is given by the shape of its three leaflets. The pig heart is a common animal model in the context of aortic valve research due to its quite similar shape compared to humans (Crick et al., 1998). The hearts were bought at a slaughterhouse, so this study is not related to any ethical issue. To extract the leaflets, the aortic root was cut out of the pig heart in a first step (see Fig. 1a). Then, the root was cut vertically in between the left-coronary and the right-coronary leaflet. Hence, the root could be opened and the three sinuses were separated using vertical incisions in between the leaflets. Finally, the leaflets were cut off of the aortic root wall. To acquire the individual shape of the extracted leaflets, they were spread on an illuminated plate and a photograph was taken. Special attention was paid to preserve the leaflets natural shape while spreading them on the plate. The illumination was monochromatic with a wavelength of 470 nm. Light of this color is strongly absorbed in the collagen fibers of the leaflets, resulting in a good contrast of the very thin structures (see Fig. 1b). Details on the setup can be found in (Hagenah et al., 2018b). This procedure was performed for 56 porcine hearts, resulting in images of 168 leaflets.

All of these images were preprocessed by applying the following pipeline. At first, the images were transformed to grayscale images and inverted. In these images, the background pixels were set to 0 using thresholding. To avoid holes in the segmentation due to very thin areas of the leaflets, the segmentation thresholds were adjusted manually for each image, ranging from 158 to 168. Afterwards, the leaflets were centered by translating the center of mass to the image's mid point and by rotating the leaflet so that the commissure points are vertically aligned. Finally, the images were downsampled to a size of  $128 \times 64$  pixels with a resolution of  $0.34 \frac{\text{mm}}{\text{pixel}}$ . The result is exemplarily shown in Fig. 1c.

## 2.2. Representation Learning

We trained an artificial neural network to find a compact yet meaningful representation of a leaflet image in an unsupervised manner. To this end, we used a convolutional autoencoder  $AE$ , which is known to be capable of encoding images with good representations in the bottleneck layer, i.e. the latent space ([Hinton and Salakhutdinov, 2006](#)). The autoencoder is divided into two different subnetworks: the encoder network  $enc : I \in \mathcal{R}^{128 \times 64} \rightarrow z \in \mathcal{R}^{n_z}$  and the decoder network  $dec : z \in \mathcal{R}^{n_z} \rightarrow I \in \mathcal{R}^{128 \times 64}$  such that

$$AE(I) = dec(enc(I)) = I_{reco}, \quad (1)$$

where  $I$  is an image of one leaflet and  $n_z$  is the dimensionality of the latent space, i.e. the number of neurons in the bottleneck layer.

One important advantage of autoencoders is that generative modeling of images based on latent representations is possible using the decoder part. Like this, the clustering can be performed in the latent space while the evaluation of the clustering can be executed in image space by propagating the latent cluster centers through the decoder. To this end, we used a symmetric autoencoder with 3 convolutional layers (ConvL) with 32 filters, each followed by a maximum-pooling layer (maxPoolL), and a fully connected layer (FCL) for the encoder. In all layers, ReLU-Activation was applied. The decoder architecture was identical but mirrored. The autoencoder was implemented in *Keras* using the *tensorflow*-backend ([Chollet et al., 2015](#)). It was trained with mean-squared-error loss using the ADAM-optimizer. An illustration of the proposed architecture can be found in the appendix in Fig. 4.

To evaluate the quality of the representation, we analyzed the reconstruction accuracy of the autoencoder by propagating the leaflet images through the encoder and generate an image of the resulting latent representation using the decoder. These predicted images were compared to their ground truth using two metrics: the Jaccard-Coefficient  $d_J$  and the Hausdorff-Metric  $d_H$  ([Yeghiazaryan and Voiculescu, 2015](#)). The Jaccard-Coeffienct measures the overlap of two leaflets and hence evaluates the overall shape similarity:

$$d_J(A, B) = \frac{A \cap B}{A \cup B}, \quad (2)$$

where  $A$  and  $B$  are the sets of non-zero pixels in the leaflet images, respectively. The Hausdorff-Metric describes the maximum of the minimal distances of two contours, measuring the detailed accurateness of the leaflet contour in the reconstruction:

$$d_H(X, Y) = \max \left( \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(y, x) \right), \quad (3)$$

where  $d$  is the euclidian distance and  $X$  and  $Y$  are sets of contour pixels, respectively. Since the reconstruction accuracy is highly dependant of the network architecture, we performed this analysis additionally for a more shallow (1 ConvL followed by 1 maxPoolL, 1 FCL) and a deeper encoder architecture (5 ConvL followed by 1 maxPoolL, respectively, 1FCL). Once again, the decoder architecture was the mirrored encoder architecture. The number of latent dimensions  $n_z$  was set to 20.

Furthermore, we evaluated the reconstruction accuracy of the autoencoder for different choices of  $n_z$ , ranging from 2 to 50 dimensions.

However, not only the reconstruction accuracy is important to ensure a sufficient representation, but also a smooth decoder function is desired. Smoothness of the decoder ensures that interpolations in the latent space provide realistic images, which is mandatory for the generation of valve images of the identified cluster centers, that will barely match one existing valve. Hence, we compared our proposed architecture to two alternative methods. First, we used the same autoencoder but introduced data augmentation (random shearing and zooming by up to 5%, translations in any direction with a maximum of 2.2mm). Second, we evaluated a variational autoencoder (VAE) of the same encoder/decoder architecture as VAEs are optimized to identify a smooth latent space implicitly (Kingma and Welling, 2013). The comparison was done quantitatively regarding the reconstruction accuracy and qualitatively using linear interpolation between two points in the latent space and propagating the interpolated points through the decoder to produce leaflet images.

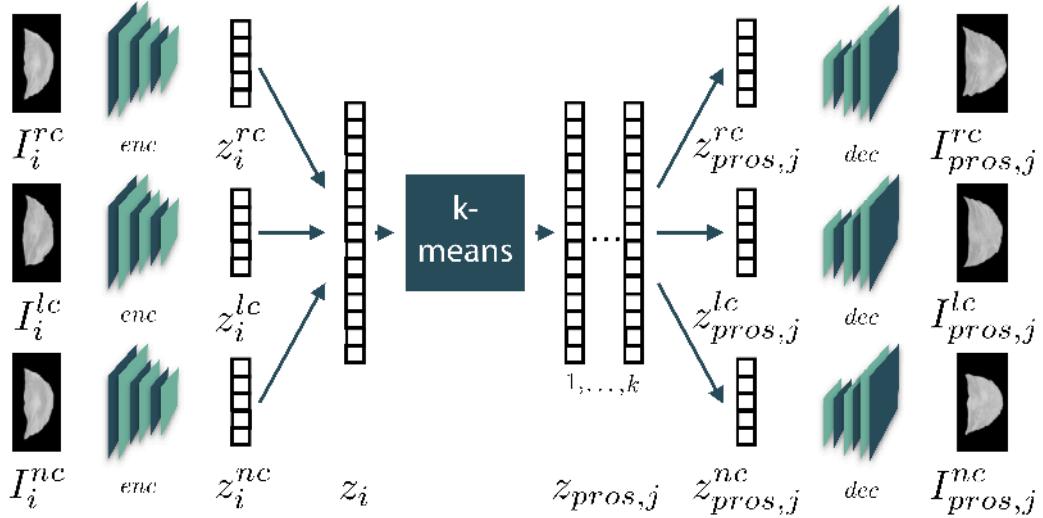


Figure 2: Sketch of the proposed method. The trained encoder model is used to receive the latent vector of each leaflet. These vectors are stitched together to represent a valve. In this space, the clustering is performed. Afterwards, the latent representation of the single leaflets of an identified prosthesis are extracted and the corresponding images are generated using the decoder.

### 2.3. Clustering

Our aim is to perform a cluster analysis in latent space. To this end, we assume that the geometry of the whole valve can be described by the shape of its three leaflets. Thus, we propagated the three leaflets of one valve through the trained encoder network to get their latent representations  $z_i^{rc} \in \mathcal{R}^{n_z}$ ,  $z_i^{lc} \in \mathcal{R}^{n_z}$  and  $z_i^{nc} \in \mathcal{R}^{n_z}$ , where  $rc$  stands for the right-coronary,  $lc$  for the left-coronary and  $nc$  for the non-coronary leaflet, respectively, and  $i = 1, \dots, 56$  describes the current valve. By stitching the latent vectors together, we retrieve a latent representation  $z_i \in \mathcal{R}^{3n_z}$  of the whole valve. In this representation, a k-means-clustering was performed, aiming on identifying  $k$  cluster centers

$z_{pros,j}, j = 1 \dots k$  that could serve as prosthesis types (Lloyd, 1982). Even though a clustering in the latent space is convenient, the evaluation of the identified prosthesis types in latent space is not intuitive. Hence, we transformed the prosthesis back to image space using the decoder. For this purpose, the latent prosthesis representation was split up into the three parts  $z_{pros,j}^{rc} \in \mathcal{R}^{n_z}$ ,  $z_{pros,j}^{lc} \in \mathcal{R}^{n_z}$  and  $z_{pros,j}^{nc} \in \mathcal{R}^{n_z}$  corresponding to the three leaflets. These vectors were propagated through the decoder, resulting in images of the three leaflets of the prosthesis, i.e.  $I_{pros,j}^{rc} \in \mathcal{R}^{128 \times 64}$ ,  $I_{pros,j}^{lc} \in \mathcal{R}^{128 \times 64}$  and  $I_{pros,j}^{nc} \in \mathcal{R}^{128 \times 64}$ . These images could be compared to the images of all valves that correspond to this cluster using the metrics described above. Hence, we can calculate the mean Jaccard coefficient  $\overline{d_{J,k}}$  and the mean Hausdorff metric  $\overline{d_{H,k}}$  over all valves in dependency of the number of identified prosthesis types  $k$  as

$$\overline{d_{J,k}} = \frac{1}{56} \sum_{j=1}^k \sum_{I_c \in C_j} \frac{1}{3} \left( d_J(I_c^{rc}, I_{pros,j}^{rc}) + d_J(I_c^{lc}, I_{pros,j}^{lc}) + d_J(I_c^{nc}, I_{pros,j}^{nc}) \right) \quad (4)$$

$$\overline{d_{H,k}} = \frac{1}{56} \sum_{j=1}^k \sum_{I_c \in C_j} \frac{1}{3} \left( d_H(I_c^{rc}, I_{pros,j}^{rc}) + d_H(I_c^{lc}, I_{pros,j}^{lc}) + d_H(I_c^{nc}, I_{pros,j}^{nc}) \right), \quad (5)$$

where  $C_j$  is the set of all images  $I_c$  corresponding to the  $j$ -th cluster. Using these metrics a set of used prosthesis types can be evaluated regarding its capability of approximating each individual valve shape in our data set. To analyze the amount of valve prosthesis types needed to ensure a good shape approximation, we performed the clustering for different values of  $k$ , ranging from 1 to 56. Note that a clustering with  $k = 1$  corresponds to the current clinical situation, where each patient is treated with the same prosthesis shape. For the clustering study,  $n_z$  was set to 20.

### 3. Results

The results of this study are divided into two parts: the analysis of the learned representation and the identification of valve types using clustering in the latent space.

#### 3.1. Representation Analysis

The results of the experiment regarding the reconstruction accuracy of our proposed architecture as well as of alternative architectures is given in Table 1. Compared to a shallower and a deeper architecture, the proposed autoencoder achieves the highest Jaccard-Coefficient and the smallest value of the Hausdorff-Metric, given as the mean over all leaflets.

Fig. 3a shows the reconstruction accuracy in dependency of  $n_z$ , the number of latent dimensions. The accuracy increases at first, but saturates at a value of about 20 dimensions. Hence, additional dimensions only lead to a slight increase of the reconstruction accuracy. The introduction of data augmentation as well as the use of a variational autoencoder provide comparable results as the proposed architecture (see Table 1). The evaluation of the smoothness of the different models was done qualitatively. While the VAE delivered slightly different shapes at the leaflet tips, no relevant differences between the proposed AE with or without data augmentation could be observed. Results are exemplarily shown in the appendix in Fig. 5.

Table 1: Reconstruction accuracy of the proposed autoencoder (marked in bold), compared to a shallower and a deeper architecture regarding the Jaccard-Coefficient  $d_J$  and the Hausdorff-Metric  $d_H$ . Additionally, the influence of data augmentation as well as the usage of a variational autoencoder is shown. The metrics are given as the mean over all leaflet images.

Architecture	$d_J$	$d_H$ [mm]
Shallower architecture	0.9272	3.17
<b>Proposed architecture</b>	<b>0.9471</b>	<b>2.60</b>
Deeper architecture	0.9283	3.18
Proposed AE with Data Augmentation	0.9424	2.70
VAE of same architecture	0.9423	2.74

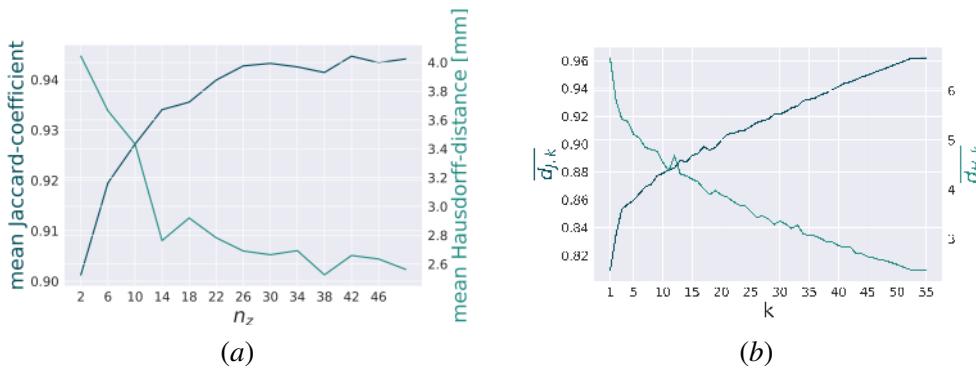


Figure 3: Resulting reconstruction accuracies. (a) Reconstruction accuracy of the autoencoder in dependency of the number of latent dimensions  $n_z$ . (b) Capability of reproducing all individual valve shapes given in the data set in dependency of the number of prosthesis type.

### 3.2. Cluster Identification

The clustering was performed for different numbers of clusters  $k$ . The capability of different numbers of prosthesis shape types to approximate the individual shapes given in our data set is shown in Fig. 3b. For low values of  $k$ , the increase of accuracy is very steep, while it starts to flatten at a value of about 7. The resulting prostheses are exemplarily shown for  $k = 3$  in the appendix in Fig. 6. To critically question the whole approach of our proposed autoencoder pipeline, we also performed the clustering and evaluation routine regarding the image description using principal components. To this end, we performed Principal Component Analysis (PCA) on the vectorized images and used the 20 most relevant components as a representation. While the mean Jaccard coefficient was significantly smaller compared to the autoencoder approach, outliers inn the range of up to 20mm were observed in the mean Hausdorff metric. Detailed results of this comparison can be found in the appendix in Fig. 7.

## 4. Discussion and Conclusion

The results show that learning a feasible representation of the aortic valve leaflets is possible using convolutional autoencoders. Due to the saturation of the metrics with  $n_z$  greater than 20, we achieve a good trade-off between compactness of the representation and reconstruction accuracy with 20 latent dimensions. Our proposed architecture delivers the best results of all compared models. It is interesting that the usage of a deeper architecture did not lead to an increase in reconstruction accuracy. This might be related to the fact that autoencoders do not encode in general and that more complex decoders are more likely to just approximate the distribution of images without being conditioned by the latent space value (Chen et al., 2016). Another possible explanation is the higher number of trainable parameters on the relatively small data set.

We could also show that no increase in reconstruction accuracy is achieved by introducing data augmentation. This might be due to the small variance in the data that is induced by the preprocessing pipeline. By accurate centering and unification, common advantages of data augmentation become irrelevant, such as translational or rotational invariance. Additionally, the general, rough shape of the leaflets is quite similar while the differences are mainly given by different contour lines or detailed shape differences, which also explains the overall high values of the Jaccard-Coefficient. In the qualitative smoothness identification, slight differences between the proposed autoencoder and the VAE are visible. Hence, the VAE derives a different latent representation. However, an improvement of the reconstruction accuracy by using the VAE could not be observed.

The cluster analysis reveals the high potential of personalization. The clustering with  $k = 1$  corresponds to the current situation where each patient is treated with the same valve geometry. By introducing more than one prosthesis type, the metrics for estimating how good the shape variance can be approximated deliver much better values. Note that the Jaccard-Coefficient for  $k = 1$  is already at 0.8. This means, that the whole potential of personalization lies within a range of a Jaccard-Coefficient increase of 0.2. Just by introducing three different prosthesis types, we can close about one quarter of this personalization gap. Likewise, the Hausdorff-distance drops by 1.25mm when three prosthesis types are used. With more than three types, the increase or decrease of the metrics is slower, but still significant. However, this is reasonable, because a data set can always be approximated better the more cluster centers are used. At a number of about  $k = 7$ , the increase appears to be approximately linear. This leads to the expected conclusion that the biggest

advantage in the trade-off between patient's outcome and economical issues can be achieved by introducing 3 – 5 prosthesis shape types.

In this study, only the geometric shape of the leaflets could be evaluated. Remaining questions are the dynamical and biomechanical improvements of the valve prosthesis' functionality achieved by the proposed personalization technique. By manufaction of the identified prostheses, a comparison between state-of-the-art prosthesis and the personalized ones could be done in a left-heart simulator to analyze the impact of personalization. Due to the manual processing of the porcine herat valves, deformations of the leaflet shapes cannot be entirely avoided. However, we assume that these errors would be normally distributed, hence, no systematic bias towards specific shapes would be present in the data set. Another limitation of the study is the transferability to the human heart. Even though the aortic valves of pigs and humans appear to be quite similar and pig valves are already used as xenological prostheses, the direct usage of the identified prosthesis shapes in humans should be further investigated. However, the proposed autoencoder model could still present a sufficient basis using transfer learning, i.e. fine-tune the autoencoder with a very small set of human valves.

To the best of our knowledge, this study presents the first approach to derive prosthesis shapes via clustering in a latent space description. Like this, the clustering focuses on abstract meta-features given by the deep representation rather than on pixel-wise differences. This makes the concept transferable to a lot of comparable problems in the area of personalized medicine. By adjusting the architecture of the autoencoder, it is possible to encode very different kinds of anatomies or images from different modalities, like 3D tomographic volumes. Hence, our method could present an important step towards personalized prosthetics.

## Acknowledgments

The authors would like to thank Tizian Evers and Meike Fünderich for their help conducting the experiments.

## References

- Talal Al-Atassi, Hadi Daood Toeg, Reza Jafar, Benjamin Sohmer, Michel Labrosse, and Munir Boodhwani. Impact of aortic annular geometry on aortic valve insufficiency: Insights from a preclinical, ex vivo, porcine model. *The Journal of thoracic and cardiovascular surgery*, 150(3):656–664, 2015.
- Tina S Andersen, Peter Johansen, Bekka O Christensen, Peter K Paulsen, Hans Nygaard, and J Michael Hasenkam. Intraoperative and postoperative evaluation of cavitation in mechanical heart valve patients. *The Annals of Thoracic Surgery*, 81(1):34–41, 2006.
- Claudia Blais, Jean G Dumesnil, Richard Baillot, Serge Simard, Daniel Doyle, and Philippe Pibarot. Impact of valve prosthesis-patient mismatch on short-term mortality after aortic valve replacement. *Circulation*, 108(8):983–988, 2003.
- Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

François Chollet et al. Keras. *GitHub*, <https://github.com/fchollet/keras>, 2015.

Simon J Crick, Mary N Sheppard, Siew Yen Ho, Lior Gebstein, and Robert H Anderson. Anatomy of the pig heart: comparisons with normal human cardiac structure. *The Journal of Anatomy*, 193(1):105–119, 1998.

Laurent De Kerchove, Mona Momeni, Gaby Aphram, Christine Watremez, Xavier Bollen, Ramadan Jashari, Munir Boodhwani, Parla Astarci, Philippe Noirhomme, and Gebrine El Khoury. Free margin length and coaptation surface area in normal tricuspid aortic valve: an anatomical study. *European Journal of Cardio-Thoracic Surgery*, 2017.

Jannis Hagenah, Tizian Evers, Michael Scharfschwerdt, Achim Schweikard, and Floris Ernst. An svr-based data-driven leaflet modeling approach for personalized aortic valve prosthesis development. *Computing in Cardiology 2018*, 2018a.

Jannis Hagenah, Michael Scharfschwerdt, and Floris Ernst. Towards personalised aortic valve prostheses: A compact description of the individual valve geometry. *Computing in Cardiology 2018*, 2018b.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Eamonn Keogh and Abdullah Mueen. Curse of dimensionality. In *Encyclopedia of machine learning*, pages 257–258. Springer, 2011.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

Philippe Pibarot and Jean G Dumesnil. Prosthetic heart valves. *Circulation*, 119(7):1034–1048, 2009.

Varduhı Yegiazaryan and Irina Voiculescu. An overview of current evaluation methods used in medical image segmentation. Technical report, Tech. Rep. CS-RR-15-08, Department of Computer Science, University of Oxford, Oxford, UK, 2015.

## Appendix A. Encoder Architecture

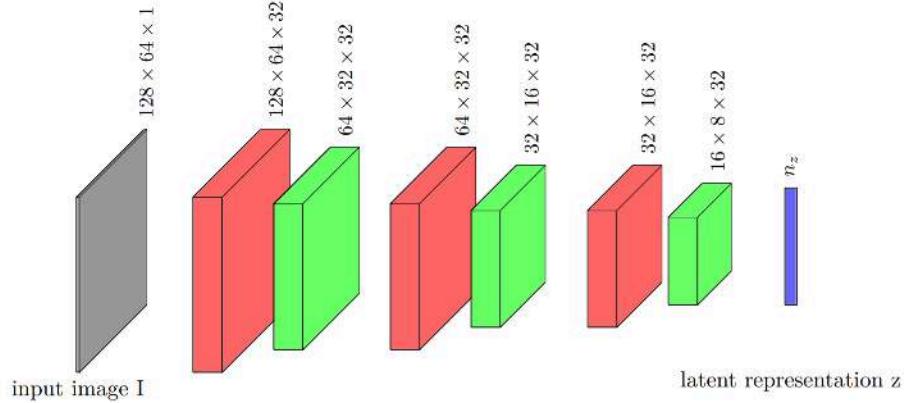


Figure 4: The architecture of the proposed encoder network, consisting of 3 convolutional layers (red), 3 maximum pooling layers (red) and a fully-connected layer (blue). The decoder consists of a similar, but mirrored architecture.

## Appendix B. Qualitative Analysis of Smoothness

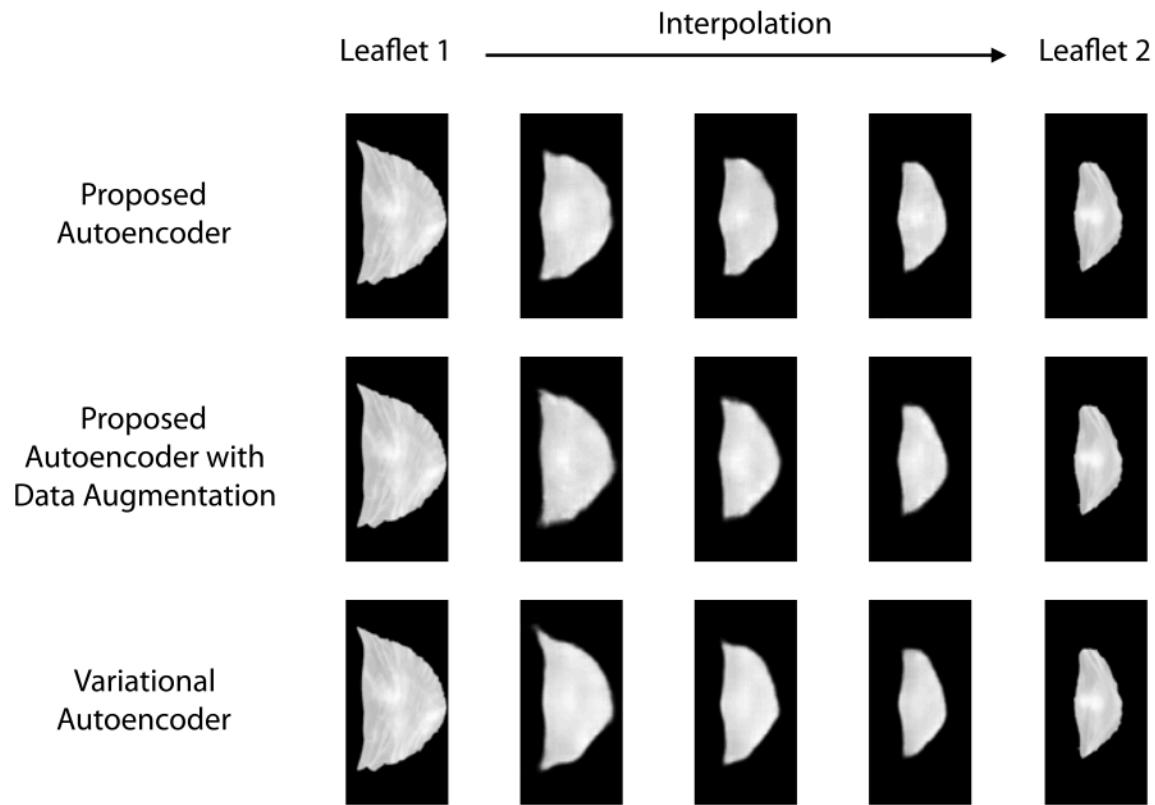
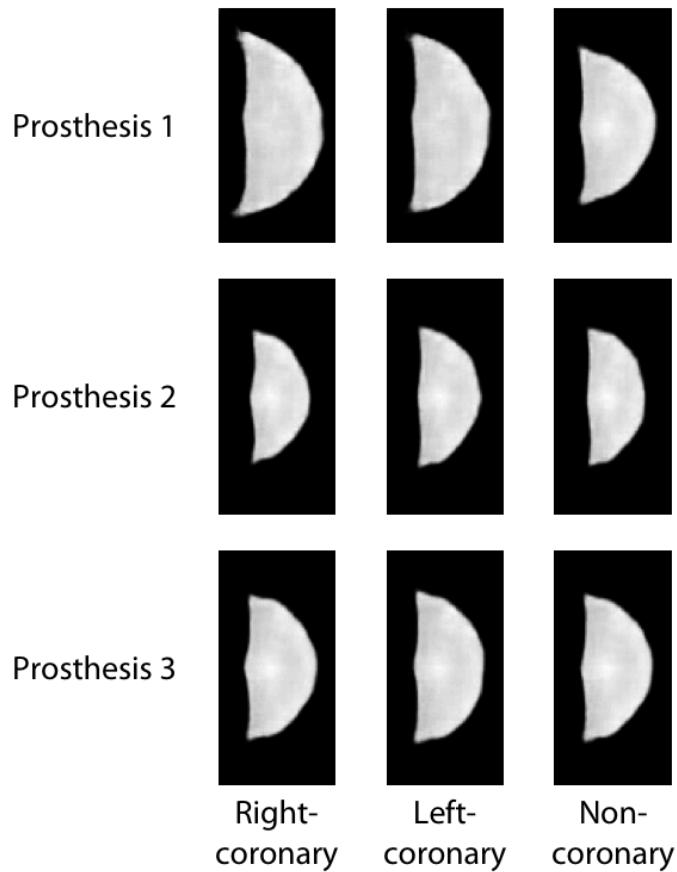


Figure 5: Smoothness of the autoencoder. Between leaflet 1 and leaflet 2, linear interpolation steps were acquired in latent space and the corresponding images were reconstructed. This was done for the proposed autoencoder without and with data augmentation as well as for the variational autoencoder.

**Appendix C. Resulting prosthesis shapes**Figure 6: Resulting prosthesis shapes, exemplarily shown for  $k = 3$ .

## Appendix D. Comparison to PCA approach

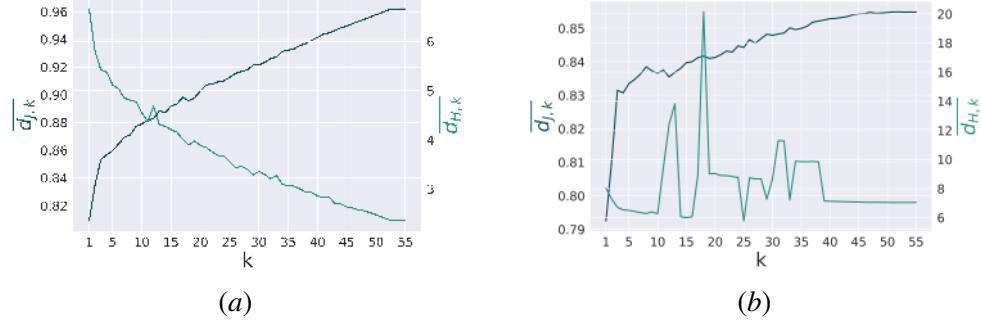


Figure 7: Capability of reproducing all individual valve shapes given in the data set in dependency of the number of prosthesis type for the proposed autoencoder (a, copied from Fig. 3b) and when Principal Component Analysis (PCA) is used for deriving the latent space representation (b). For comparison, the number of principal components was set to  $n_z = 20$ .

# Sparse Structured Prediction for Semantic Edge Detection in Medical Images

**Lasse Hansen**

**Mattias P. Heinrich**

*Institute of Medical Informatics, University of Lübeck, DE*

HANSEN@IMI.UNI-LUEBECK.DE

HEINRICH@IMI.UNI-LUEBECK.DE

## Abstract

In medical image analysis most state-of-the-art methods rely on deep neural networks with learned convolutional filters. For pixel-level tasks, e.g. multi-class segmentation, approaches build upon UNet-like encoder-decoder architectures show impressive results. However, at the same time, grid-based models often process images unnecessarily dense introducing large time and memory requirements. Therefore it is still a challenging problem to deploy recent methods in the clinical setting. Evaluating images on only a limited number of locations has the potential to overcome those limitations and may also enable the acquisition of medical images using adaptive sparse sampling, which could substantially reduce scan times and radiation doses.

In this work we investigate the problem of semantic edge detection in CT and X-ray images from sparse sampling locations. We propose a deep learning architecture that comprises of two parts: 1) a lightweight fully convolutional CNN to extract informative sampling points and 2) our novel sparse structured prediction network (SSPNet). The SSPNet processes image patches on a graph generated from the sampled locations and outputs semantic edge activations for each patch which are accumulated in an array via a weighted voting scheme to recover a dense prediction. We conduct several ablation experiments for our network on a dataset consisting of 10 abdominal CT slices from VISCERAL and evaluate its performance against strong baseline UNets on the JSRT database of chest X-rays.

**Keywords:** sparsity, structured prediction, edge detection, deep learning.

## 1. Introduction

The vast majority of medical image acquisition and analysis has so far focused on reconstructing and processing dense data. This is mainly motivated by the simplicity of representing data points and their spatial relationships on regular grids and storing or visualizing them using arrays. In particular convolutional operators for feature extraction and pooling have seen increased importance for denoising, segmentation, registration and detection due to the rise of deep learning techniques. Learning spatial filter coefficients through backpropagation is well understood and computationally efficient due to highly optimized matrix multiplication routines for both CPUs and GPUs. However, for many computer vision tasks in medical image analysis such as landmark or edge detection it seems unnecessary and expensive (in terms of time and memory limitations) to process images end-to-end with dense methods, e.g. fully-convolutional networks or encoder-decoder architectures. Therefore, in this work, we aim to show new possibilities in the area of deep learning to process image data on sparse and irregular instead of dense grids. The feasibility of our suggested approach is demonstrated on the problem of semantic edge detection in CT and X-ray images.

**Related Work:** Of all hierarchical feature learning models, CNNs have shown to be one of the most successful approaches for a wide variety of tasks such as classification, bounding box regression and segmentation (Ronneberger et al., 2015; He et al., 2017). Lately, another class of works (graph convolutional neural networks (GCNNs)) attempts to transfer these well-known concepts from the two dimensional image domain to non-Euclidean and irregular domains. Spectral CNNs, defined on graphs, were first introduced in (Bruna et al., 2013). The main drawback of the proposed method is that it relies on prior knowledge of the graph structure to define a local neighborhood for weight sharing. (Henaff et al., 2015) extended the ideas to graphs where no prior information on the structure is available. While (Bruna et al., 2013; Henaff et al., 2015) relied on splines for the formulation of their graph convolutional operators, (Kipf and Welling, 2017) uses truncated Chebyshev polynomials that allow for clear description of the support size of the learned spectral filters. (Bronstein et al., 2017) provides a comprehensive review of current research on this topic. In the medical domain GCNNs were successfully applied in a number of applications such as population-based disease prediction (Parisot et al., 2017), metric learning for brain connectivity graphs (Ktena et al., 2017) and survival analysis on pathological images (Li et al., 2018).

Edge detection is a key task in computer vision applications and is studied for decades (Canny, 1986). (Dollár and Zitnick, 2013) chose a data-driven approach using random decision forests to predict structured labels from input image patches. This technique was successfully applied in the medical domain for multi-modal registration of ultrasound and CT/MRI images (Oktay et al., 2015). (Xie and Tu, 2015) is the first deep learning method to explicitly learn edges. Features are extracted with a modified VGGNet and all layers are trained with deep supervision. In the end side outputs from different VGG Layers are fused to output a final edge map. State-of-the-art detectors for semantic edge detection mainly resemble encoder-decoder architectures that are trained with specialized loss terms (Yu et al., 2017; Liu et al., 2018).

**Contributions:** In this work we make a first step towards dense prediction from a few sparse sampling points using deep learning methods. We bring together the robustness of grid based CNNs and the flexibility of GCNNs in a single framework for pixel-level structured prediction. In this, our work differs from (Li et al., 2018), which used GCNNs for global context aggregation for image labeling. Our main contributions is the sparse structured prediction network (SSPNet). Furthermore, we successfully provide a first proof-of-concept for our new approach by evaluating it on the challenging task of semantic edge detection in medical images.

## 2. Methods

In this section, we present our proposed approach for sparse structured prediction for semantic edge detection. Figure 1 illustrates the general idea of our method. Input to our pipeline is an image  $x$ . A light-weight CNN  $\phi$ , called sample CNN, extracts potentially informative locations from the image and outputs a single channel sample map  $\phi(x)$ . A fixed number  $N$  of sample coordinates  $((x_1, y_1), \dots, (x_N, y_N))$  are drawn following a multinomial distribution with probabilities proportional to the sample map's values. Depending on the application many alternatives of extracting sampling locations are conceivable, e.g. for landmark detection one could initialize the sample map with the mean locations of the landmarks in the training set. At the given positions, patches  $(p_1, \dots, p_N)$  are extracted from the input image  $x$ . Furthermore, a simple distance graph  $G_\sigma$

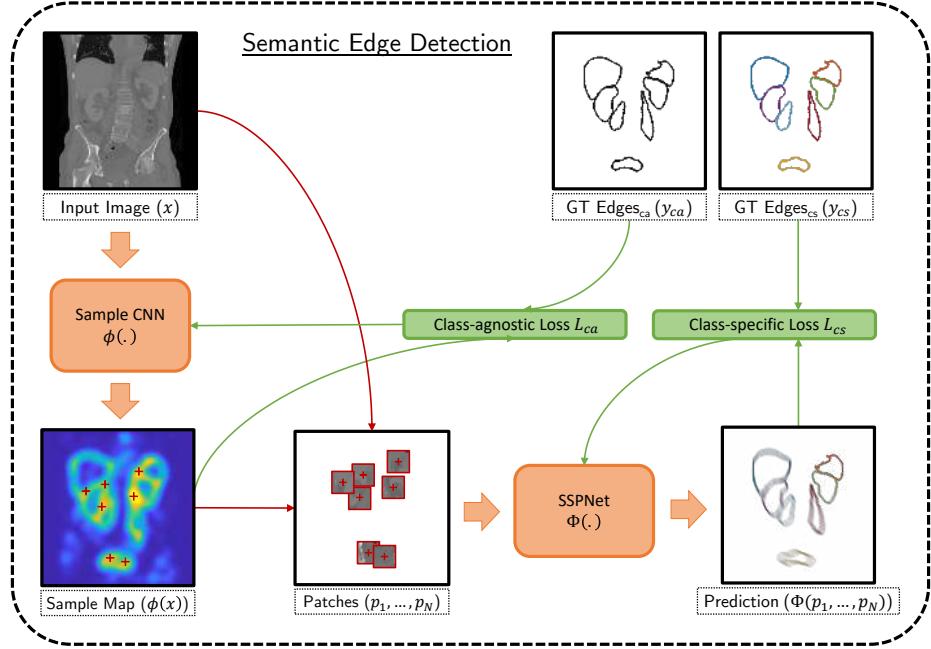


Figure 1: Our general idea for sparse semantic edge detection. We train a lightweight fully-convolutional CNN with a class-agnostic loss to output an informative heatmap from which samples are drawn with probabilities proportional to its values. Image patches are extracted around the chosen locations and our proposed SSPNet processes the generated patch graph to output a semantic edge activation for each sampling point. To recover a dense prediction all edges are accumulated in an array and the class-specific loss is applied to update the SSPNet’s parameters.

is generated. The adjacency matrix  $\mathbf{A}$  of the graph  $G_\sigma$  is given by entries

$$a_{ij} = \exp\left(\frac{-d_{ij}^2}{2 \cdot \sigma^2}\right),$$

where  $\sigma$  is a scalar diffusion coefficient and  $d_{ij}$  denotes the euclidean distance between two sampling locations  $(x_i, y_i)$  and  $(x_j, y_j)$ . Again, depending on the application and given priors the graph may be initialized accordingly. Next, the extracted image patches  $(p_1, \dots, p_N)$  as well as the graph  $G_\sigma$  serve as input to our proposed SSPNet (explained in detail below), which predicts edges for each input image patch and accumulates all predictions on a dense grid weighted by their class-specific confidence. This semantic edge map is our final output. While the focus of this work is clearly on the SSPNet, in the following we also shortly describe the training of the sample CNN.

## 2.1. Sample CNN

The sample CNN  $\phi$  is based on a lightweight version of the holistically-nested architecture in (Xie and Tu, 2015). We significantly cut the networks capacity by removing deeper layers and use

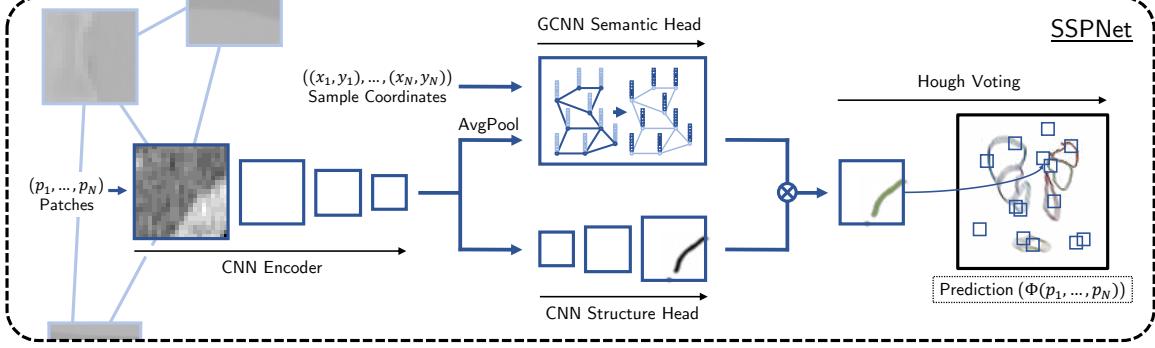


Figure 2: The proposed sparse structured prediction net (SSPNet) expects a graph of patches sampled at informative image locations. For each patch a CNN encoder extracts a set of feature maps, which are further processed by 1) the structure head that predicts local edge activations and 2) the semantic head where global context is aggregated by a GCNN. We perform a weighted Hough voting to accumulate all predictions and recover a dense semantic edge map.

reduced numbers of filters. In total the network consists of only three layers (each with two  $3 \times 3$  convolutions + relu activation). Layer 2 and 3 start with convolutions with stride 2 resulting in the network’s receptive field size of 23. After each layer a side output  $\hat{y}_i$  is generated by a further  $1 \times 1$  convolution and sigmoid activation. Side outputs are concatenated and fused to form a final prediction  $\hat{y}_0$  by a  $1 \times 1$  convolution and sigmoid activation. The sample CNN is trained with deep supervision on all outputs using the loss function from (Deng et al., 2018) which combines the binary cross-entropy (*BCE*) and Dice loss (*DICE*) to

$$L_{ca} = \sum_{i=0}^3 \alpha BCE(\hat{y}_0, y_{ca}) + \beta DICE(\hat{y}_0, y_{ca}).$$

In the loss term  $y_{ca}$  depicts the class-agnostic version of the ground truth edge map and  $\alpha$  and  $\beta$  control the weighting of the two losses. Trading robustness for precise localization the final sample map is obtained from prediction  $y_0$  after multiple average pooling steps with stride 1.

## 2.2. SSPNet

As stated above input for our SSPNet  $\Phi$  are the extracted image patches  $(p_1, \dots, p_N)$  as well as the graph  $G_\sigma$ . The network itself consists of a CNN encoder part, a structure and semantic head and a final Hough voting step to recover a dense prediction from the single patches. An overview of our proposed SSPNet is given in Figure 2. The CNN encoder applies four convolutions (kernel sizes: 5, 3, 3, 3) with relu activations on each image patch. The resulting feature maps are further processed by two network heads. The structure head consists of three transposed convolutions (kernel sizes 3, 3, 3) and relu activations. The final structured prediction is obtained by a  $1 \times 1$  convolution with sigmoid activation. The semantic head aggregates global context information and is modeled with a GCNN. Input features on our graph are the average pooled feature maps from the CNN encoder. We

also experiment with explicitly adding the sampling coordinates as additional informative features. For this part of our work we decided to strive for simplicity and use a simple random walk diffusion with a single  $\sigma$  kernel to pool features across our input graph  $G_\sigma$  (Atwood and Towsley, 2016; Hansen et al., 2018). The diffusion process can be described by the diffusion matrix

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A},$$

where  $\mathbf{I}$  denotes the identity matrix and the degree matrix  $\mathbf{D}$  is solely defined by its diagonal elements  $d_{ii} = \sum_j a_{ij}$ . By matrix multiplication with the input feature vector a weighted average pooling across edges of the graph is employed. The diffusion pooling is followed by two  $1 \times 1$  convolutions with relu activations. In total we employ two of the described graph convolutions. Final semantic confidence scores for each node (sampling locations) on the graph are obtained by a  $1 \times 1$  convolution with sigmoid activation. As Hough voting has been proven to be effective for locating shapes in images (Ballard, 1981; Lindner et al., 2015) we accumulate the structured predictions from all image patches on a dense grid (with  $C$  channels, where  $C$  corresponds to the number of semantic classes) and weight each prediction with the corresponding semantic confidence score. Furthermore, each grid point is normalized by the number of predictions made for this point. Note that by construction the final semantic edge map holds values between 0 and 1 and we can apply our class-specific similar to our class-agnostic loss as

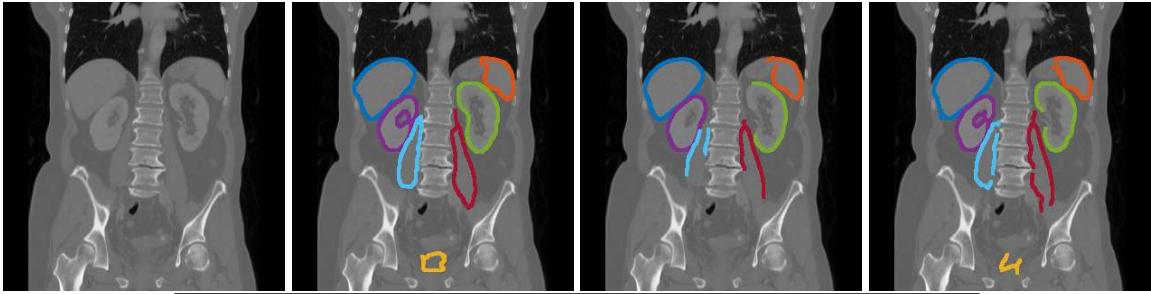
$$L_{cs} = \sum_{i=0}^{C-1} w_i (\alpha BCE(\hat{y}^{(i)}, y_{cs}^{(i)}) + \beta DICE(\hat{y}^{(i)}, y_{cs}^{(i)})),$$

where  $y_{cs}$  depicts the one-hot encoded semantic ground truth edges, such that pixels can belong to multiple labels. Classes may be weighted by the parameters  $w_i$ .

### 3. Experiments and Results

We validate the feasibility of our approach on two different datasets for the task of semantic edge detection. The first dataset consists of 10 2D coronal slices of abdominal CT scans from VISCERAL (Jimenez-del Toro et al., 2016) and the second dataset is the JSRT database of 247 chest X-ray images (Shiraishi et al., 2000). As validation metric we use the F-score on the fixed contour threshold (ODS), where the threshold is determined from all images in the test dataset. Before evaluation the thresholded predictions are thinned and spurious detections ( $< 10$  pixels) are removed. ODS metrics are computed for each semantic class individually and we report the mean value. We compare our approach against three different 5-Layer UNet implementations (UNet-S, UNet-M, UNet-L). The UNet-S has a comparable capacity in terms of learnable parameters as our SSPNet, whereas the UNet-L has almost 2.5 as many parameters.

**Implementation Details:** All models were trained for 300 and 100 epochs for VISCERAL and JSRT, respectively. ADAM optimization was used with an initial learning rate of .02. We employ batch normalization with a mini batch size of 4 and an exponential learning rate schedule with a multiplicative factor of 0.99 to stabilize training. The images are augmented with a random affine transformation. The graph for the SSPNet is computed with a  $\sigma$  value of 0.1 and normalized coordinates. We set the  $\alpha$  and  $\beta$  parameters of the loss terms to .001 and 1, respectively. Class weights were applied corresponding to the organ label occurrences for all experiments with the UNet variants. During training and at test time we sample patches at 500 and 2000 locations, respectively. All hyperparameters were determined by grid search for our simplest baseline method and kept fixed for all further experiments.



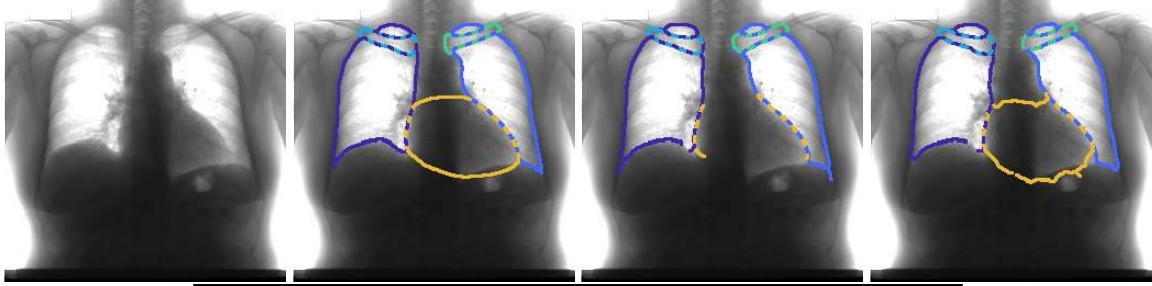
Method	Parameters	Samples	ODS
SSPNet – $1 \times 1$ conv.	~ 500k	2000	.690
SSPNet – GCNN	~ 500k	2000	.786
SSPNet – $1 \times 1$ conv. + coords.	~ 500k	2000	.801
<b>SSPNet – GCNN + coords.</b>	~ 500k	2000	<b>.827</b>
UNet-S	~ 500k	dense	.769
UNet-M	~ 900k	dense	.791
<b>UNet-L</b>	~ 1300k	dense	<b>.834</b>

Figure 3: Qualitative and quantitative results on VISCERAL. The images (from left to right: original CT slice, ground truth, UNet-L, SSPNet) show edge overlays from seven anatomical structures: liver ■, spleen □, bladder ▨, left kidney ▤, right kidney ▢, left psoas major muscle (pmm) ▲ and right pmm △. Our approach outlines edges of the psoas muscles much clearer and also detects the urinary bladder.

### 3.1. VISCERAL

We perform initial experiments on the 10 2D coronal slices of abdominal CT scans from VISCERAL in a leave-one-out fashion. The images are resampled to an isotropic pixelsize of  $1.5\text{mm}^2$  and cropped to dimensions of  $320 \times 312$  without any guidance. We consider ground truth labels for seven anatomical structures: liver ■, spleen □, bladder ▨, left kidney ▤, right kidney ▢, left psoas major muscle (pmm) ▲ and right pmm △. Besides our described architecture we test three other baselines of the approach: A SSPNet employing only  $1 \times 1$  convolutions instead of the GCNN, the GCNN without sampling coordinates as additional input features and the network of  $1 \times 1$  convolutions with sampling coordinates as additional features.

**Results:** Qualitative and quantitative results are depicted in Figure 3. The GCNN outperforms the network with only  $1 \times 1$  convolutions in both cases with and without sampling coordinates as additional features, although the result is much clearer in the second case (ODS of .690 against .786). The best SSPNet with an ODS of .827 yields a higher score than the UNet-S and Unet-M and performs only slightly worse than the UNet-L (ODS of .769, .791 and .834 respectively). Without class-weighting the UNet variants perform worse with ODS values of .763, .773 and .817, respectively. In contrast, the SSPNet showed similar results with and without class weighting. The visual comparison shows a clearer outline of the psosas muscles and a better detection of the unary bladder in favor of the SSPNet.



Method	Parameters	Samples	ODS
<b>SSPNet – GCNN + coords.</b>	~ 500k	2000	.900
UNet-S	~ 500k	dense	.874
UNet-M	~ 900k	dense	.878
<b>UNet-L</b>	~ 1300k	dense	.884

Figure 4: Qualitative and quantitative results on the JSRT chest X-ray database. The images (from left to right: original X-ray, ground truth, Unet-L, SSPNet) show edge overlays from five anatomical structures: left lung ■, right lung ■, left clavicle ■, right clavicle ■ and heart ■. The UNet misses parts of the edges of the heart whereas our approach successfully follows informative gradients along its outline.

### 3.2. JSRT

The JSRT database consists of 247 chest X-ray images that were downsampled to dimensions of  $256 \times 256$ . A four-fold cross validation was employed to compute the results. We test the SSPNet with additional sampling coordinates as input features against the three UNet implementations UNet-S, UNet-M and UNet-L. Ground truth labels are generated from the provided landmarks for five anatomical structures: left lung ■, right lung ■, left clavicle ■, right clavicle ■ and heart ■

**Results:** Qualitative and quantitative results are depicted in Figure 4. The SSPNet yields a slightly higher OSD score of .900 than the UNet-L with .884, though visual results are mostly comparable. However, in some cases the UNet misses parts of the edges of the heart whereas the SSPNet can follow informative gradients along its outline.

## 4. Discussion and Conclusion

In this work we proposed a new approach for structured prediction for semantic edge detection from a few sparse sampling locations on an image. To the best of our knowledge the SSPNet is the first deep learning network that combines structured prediction with CNNs and global context aggregation with graph convolutions to recover a dense output. In our experiments on VISCERAL and JSRT we showed that the SSPNet performed better or on par with several UNet variants while also having the lowest number of trainable parameters.

For future work, incorporating the SSPNet in an end-to-end learning framework instead of working with an explicitly trained sample CNN is clearly of high interest. This may be achieved by using a more complex GCNN model with attention mechanisms, e.g. (Monti et al., 2018), which could lead the selection of sampling locations. With an extension to 3D volumes, our approach can be evaluated on medical datasets with stronger memory and computational limitations. While in this work the focus was on edge detection, other tasks for structured prediction, such as landmark detection in medical images, may also be suited well for our approach.

In conclusion, we showed that our SSPNet is feasible for semantic edge detection in medical images and we believe that it can be used as a potential alternative to dense encoder-decoder architectures for general pixel-level image tasks in deep learning.

## Acknowledgments

We would like to thank the reviewers for their many insightful comments and suggestions helping to improve our paper. We gratefully acknowledge the support of the NVIDIA Corporation with their GPU donations for this research.

## References

- James Atwood and Don Towsley. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1993–2001, 2016.
- Dana H Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.
- Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *Signal Processing Magazine*, 34(4):18–42, 2017.
- Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR)*, pages 1–14, 2013.
- John Canny. A computational approach to edge detection. *Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.
- Ruoxi Deng, Chunhua Shen, Shengjun Liu, Huibing Wang, and Xinru Liu. Learning to predict crisp boundaries. In *European Conference on Computer Vision (ECCV)*, pages 562–578, 2018.
- Piotr Dollár and C Lawrence Zitnick. Structured forests for fast edge detection. In *International Conference on Computer Vision (CVPR)*, pages 1841–1848, 2013.
- Lasse Hansen, Jasper Diesel, and Mattias P Heinrich. Multi-kernel diffusion cnns for graph-based learning on point clouds. *arXiv preprint arXiv:1809.05370*, 2018.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

- Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.
- Oscar Jimenez-del Toro, Henning Müller, Markus Krenn, Katharina Gruenberg, Abdel Aziz Taha, Marianne Winterstein, Ivan Eggel, Antonio Foncubierta-Rodríguez, Orcun Goksel, András Jakab, et al. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. *Transactions on Medical Imaging*, 35(11):2459–2475, 2016.
- Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.
- Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert. Distance metric learning using graph convolutional networks: Application to functional brain networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 469–477, 2017.
- Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 174–182, 2018.
- C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes. Robust and accurate shape model matching using random forest regression-voting. *Transactions on Pattern Analysis & Machine Intelligence*, 37(9):1862–1874, 2015.
- Yun Liu, Ming-Ming Cheng, JiaWang Bian, Le Zhang, Peng-Tao Jiang, and Yang Cao. Semantic edge detection with diverse deep supervision. *arXiv preprint arXiv:1804.02864*, 2018.
- Federico Monti, Oleksandr Shchur, Aleksandar Bojchevski, Or Litany, Stephan Günnemann, and Michael M Bronstein. Dual-primal graph convolutional networks. *arXiv preprint arXiv:1806.00770*, 2018.
- Ozan Oktay, Andreas Schuh, Martin Rajchl, Kevin Keraudren, Alberto Gomez, Mattias P Heinrich, Graeme Penney, and Daniel Rueckert. Structured decision forests for multi-modal ultrasound image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 363–371, 2015.
- Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero Moreno, Ben Glocker, and Daniel Rueckert. Spectral graph convolutions for population-based disease prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 177–185, 2017.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.
- Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development

of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *International Conference on Computer Vision (CVPR)*, pages 1395–1403, 2015.

Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam. Casenet: Deep category-aware semantic edge detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–26, 2017.

# Exclusive Independent Probability Estimation using Deep 3D Fully Convolutional DenseNets: Application to IsoIntense Infant Brain MRI Segmentation

**Seyed Raein Hashemi<sup>1,2</sup>**

**Sanjay P. Prabhu<sup>1</sup>**

**Simon K. Warfield<sup>1</sup>**

**Ali Gholipour<sup>1</sup>**

<sup>1</sup> Computational Radiology Laboratory, Boston Children’s Hospital; and Harvard Medical School

<sup>2</sup> College of Computer and Information Science, Northeastern University; Boston, MA

## Abstract

The most recent fast and accurate image segmentation methods are built upon fully convolutional deep neural networks. In particular, densely connected convolutional neural networks (DenseNets) have shown excellent performance in detection and segmentation tasks. In this paper, we propose new deep learning strategies for DenseNets to improve segmenting images with subtle differences in intensity values and features. In particular, we aim to segment brain tissue on infant brain MRI at about 6 months of age where white matter and gray matter of the developing brain show similar T1 and T2 relaxation times, thus appear to have similar intensity values on both T1- and T2-weighted MRI scans. Brain tissue segmentation at this age is, therefore, very challenging. To this end, we propose an exclusive multi-label training strategy to segment the mutually exclusive brain tissues with similarity loss functions that automatically balance the training based on class prevalence. Using our proposed training strategy based on similarity loss functions and patch prediction fusion we decrease the number of parameters in the network, reduce the number of training classes, focusing the attention on less number of tasks, while mitigating the effects of data imbalance between labels and inaccuracies near patch borders. By taking advantage of these strategies we were able to perform fast image segmentation (less than 90 seconds per 3D volume) using a network with less parameters than many state-of-the-art networks (1.4 million parameters), overcoming issues such as 3D vs 2D training and large vs small patch size selection, while achieving the top performance in segmenting brain tissue among all methods tested in first and second round submissions of the isointense infant brain MRI segmentation (iSeg) challenge according to the official challenge test results. Our strategy improved the training process through balanced training and reduced complexity, and provided a trained model that works for any size input image and is faster and more accurate than many state-of-the-art methods.

**Keywords:** Deep learning, Convolutional Neural Network, FC-DenseNet, Segmentation

## 1. Introduction

Deep convolutional neural networks have shown great potential in medical imaging on account of dominance over traditional methods in applications such as segmentation of neuroanatomy (Bui et al., 2017; Moeskops et al., 2016; Zhang et al., 2015; Chen et al., 2017), lesions (Valverde et al., 2017; Brosch et al., 2015; Kamnitsas et al., 2017; Hashemi et al., 2019), and tumors (Havaei et al., 2017; Pereira et al., 2016; Wachinger et al., 2017) using voxelwise networks (Moeskops et al., 2016;

Havaei et al., 2017; Salehi et al., 2017, 2018), 3D voxelwise networks (Chen et al., 2017; Kamnitsas et al., 2017) and Fully Convolutional Networks (FCNs) (Çiçek et al., 2016; Milletari et al., 2016; Salehi et al., 2017, 2018; Hashemi et al., 2019). FCNs have shown better performance while also being faster in training and testing than voxelwise methods (Salehi et al., 2017, 2018).

Among these, the densely connected networks, referred to as DenseNets (Huang et al., 2017) and a few of its extensions, such as a 3D version called DenseSeg (Bui et al., 2017) and a fully convolutional two-path edition (FC-DenseNet) (Jegou et al., 2017), have shown promising results in image segmentation tasks (Dolz et al., 2018). For example the DenseSeg showed top performance in the 2017 MICCAI isointense infant brain MRI segmentation (iSeg) grand challenge (Wang et al., 2019)<sup>1</sup>, which is considered a very difficult image segmentation task for both traditional and deep learning approaches. During early infant brain maturation through the myelination process, there is an isointense period in which the T1 and T2 relaxation times of the white matter (WM) and gray matter (GM) tissue become similar, resulting in isointense (similar intensity) appearance of tissue on both T1-weighted and T2-weighted MRI contrasts. This happens around 6 months of age where tissue segmentation methods that are based directly on image intensity are prone to fail (Wang et al., 2013). Deep learning methods, however, have shown promising results in this application. In this work, we aimed to further improve image segmentation under these challenging conditions. While the top performing methods in the iSeg challenge relied on the power of DenseNets and used conventional training strategies based on cross-entropy loss function (Bui et al., 2017; Dolz et al., 2018), in this work we focused on the training part and developed new strategies that helped us achieve the best performance currently reported on the iSeg challenge among all first<sup>2</sup> and second round submissions<sup>3</sup>. We built our technique over a deep 3D two-channel fully convolutional DenseNet; and trained it purposefully using our proposed exclusive multi-label multi-class method of training, with exclusively adjusted similarity loss functions on large overlapping 3D image patches. We overcame class similarity issues by focusing the training on one of the isointense class labels (WM) instead of both (thus referred to as exclusive multi-label multi-class), where we balanced precision and recall separately for each class using  $F_\beta$  loss functions (Hashemi et al., 2019) with  $\beta$  values adjusted with respect to class prevalence in the training set.

Our contributions that led to improved iso-intense image segmentation include 1) An exclusive multi-label multi-class training approach (through independent probability estimation) using automatically-adjusted similarity loss functions per class; 2) utilizing a 3D FC-DenseNet architecture adopted from (Jegou et al., 2017) that is deeper, has more total number of skip connections and has less overall trainable parameters than networks in previous studies; and 3) training and testing on large overlapping 3D image patches with a patch prediction fusion strategy (Hashemi et al., 2019) that enabled intrinsic data augmentation and improved segmentation in patch borders while having the advantage of using any size image. Similarity loss functions, such as the Dice similarity loss, were previously proposed for two-class segmentation in V-Net (Milletari et al., 2016). The  $F_\beta$  loss functions, which showed excellent performance in training deep networks for highly imbalanced medical image segmentation (Hashemi et al., 2019), appeared to be effective also in exclusive multi-label training of DenseNets for independent multi-class segmentation in this work, where the class imbalance hyper-parameter  $\beta$  was directly adjusted based on training data in the training phase.

---

1. <http://iseg2017.web.unc.edu/>

2. <http://iseg2017.web.unc.edu/rules/results>

3. <http://iseg2017.web.unc.edu/evaluation-on-the-second-round-submission>

The official results on iSeg test data show that our method outperformed all previously published and reported methods improving DenseNets while standing in the first place after the second round submissions as of December 2018. Our proposed training strategy can be extended to other applications for independent multi-class segmentation and detection with multiple very similar and unbalanced classes. After a brief technical description of the isointense infant brain MRI segmentation challenge in the Motivation, the details of the methods, including the network architecture and our proposed strategies for training are presented in the Methods section; and are followed by the Results and Conclusions.

## 2. Motivation

The publicly available MICCAI grand challenge on 6-month infant brain MRI segmentation (iSeg) dataset contains pre-processed T1- and T2-weighted MR images of 10 infant subjects with manual labels of cerebrospinal fluid (CSF), WM, and GM for training and 13 infant subjects without labels for testing which are all pre-defined by challenge officials. The intensity distribution of all classes (CSF, GM, WM) is shown in Figure S1 in Appendix C, which shows that the intensity values of GM and WM classes on both MRI scans largely overlap. The GM-WM isointense appearance only happens around this stage of brain maturation and hinders GM-WM segmentation. CSF, which shows less overlap with GM and WM, shows a relatively spread intensity distribution, which is partly attributed to partial voluming effects in relatively large voxels where signal is averaged in CSF-GM interface, and inclusion of some other tissues such as blood vessels in the CSF label in iSeg data.

In the iSeg training data, the number of voxels in each class label is different and can be roughly presented as the average ratio of (36, 1, 2, 1.5) for non-brain, CSF, GM, and WM classes, respectively. Unbalanced labels can make the training process converge to local minima resulting in sub optimal performance. The predictions, thus, may lean towards the GM class especially when distinguishing between the isointense areas of GM and WM. Using our proposed exclusive multi-label multi-class training method, which can be extended to other segmentation or detection tasks with very similar (isointense) while exclusive labels (each voxel belonging to a single label), we aimed to 1) let the network focus on and learn one of the segmentation challenges at a time rather than two (in this case WM rather than both GM and WM), 2) reduce the bias in training towards classes with higher prevalence (in this case GM), and 3) use dedicated impartial asymmetric similarity loss functions on each of the non-similar classes independently (in this case WM and CSF).

## 3. Methods

### 3.1. Network architecture

In traditional densely connected networks each layer is connected to every other layer to preserve both high- and low-level features, in addition to allowing the gradients to flow from bottom layers to top layers resulting in more accurate predictions. Unlike Resnets (He et al., 2016) which only sum the output of the identity function at each layer with a skip connection from the previous layer, DenseNets (Huang et al., 2017) significantly improve the flow of information throughout the network by 1) using concatenation instead of summation and 2) forward connections from all

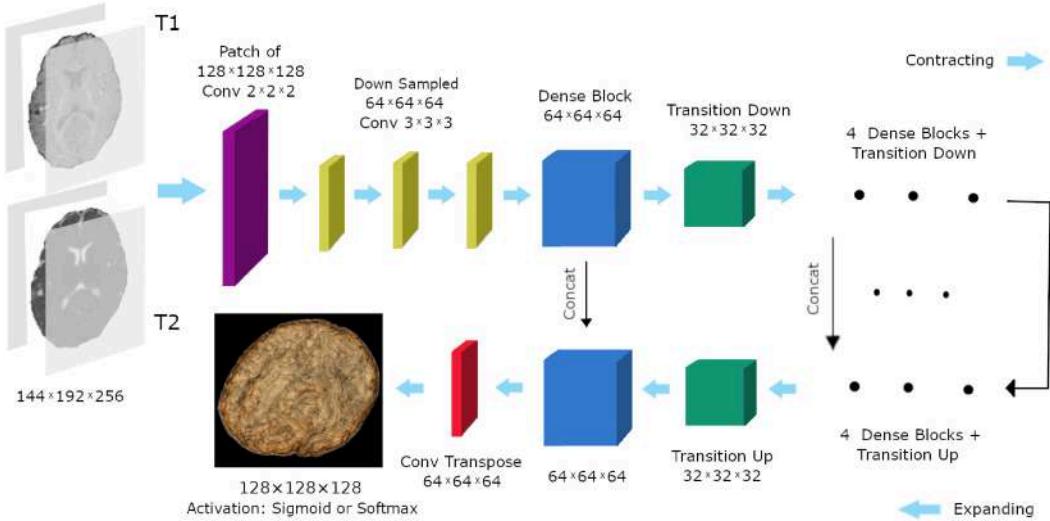


Figure 1: The 3D FC-DenseNet architecture used in this study; In the first layer, the input patch is downsampled from  $128 \times 128 \times 128$  to  $64 \times 64 \times 64$  using a  $2 \times 2 \times 2$  convolution with stride 2 (purple). Before the activation layer the patch is upsampled from  $64 \times 64 \times 64$  to  $128 \times 128 \times 128$  using a  $2 \times 2 \times 2$  convolution transpose with stride 2 (red). Using this deep architecture, we mitigated memory size limitations with large input patches, while maintaining a large field of view and incorporating 5 skip connections to improve the flow of local and global feature information.

preceding layers rather than just a previous layer, therefore:

$$x_l^{(Resnet)} = H_l(x_{l-1}) + x_{l-1} \quad , \quad x_l^{(Densenet)} = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (1)$$

where  $x_l$  is the output of the  $l^{th}$  layer,  $H_l$  is the  $l^{th}$  layer transition, and  $[x_0, x_1, \dots, x_{l-1}]$  refers to the concatenation of all previous feature maps.

We designed our deep 3D densely connected network based on a combination of DenseSeg (Bui et al., 2017) and FC-DenseNet (Jegou et al., 2017) architectures. This deep DenseNet (Huang et al., 2017) style architecture is shown in Figure 1 consisting of contracting and expanding paths. The network is trained on local features in the contracting path concatenated with upsampled global features in the expanding path. For this reason, the model has the capacity to learn both high-resolution local and low-resolution global 3D features. The depth of the architecture as well as the 5 skip-layer connections ensure the use of 5 various resolutions of local and global 3D features in the final prediction.

The contracting path contains an initial  $2 \times 2 \times 2$  convolution with stride 2 for the purpose of downsampling and reducing the patch size (to  $64 \times 64 \times 64$ ) while preserving the larger field of view ( $128 \times 128 \times 128$ ). It is then followed by three padded  $3 \times 3 \times 3$  convolutional layers. Five dense blocks follow with a growth rate of 12. Growth rate for dense blocks is the increase amount in the number of feature maps after each block. Dense blocks contain four  $3 \times 3 \times 3$  convolutional layers preceding with  $1 \times 1 \times 1$  convolutions which are known as bottlenecks (Huang et al., 2017).

Dimension reduction of 0.5 applied at transition layers helps parameter efficiency and reduce the number of input feature maps. There are skip connections between each and every layer inside dense blocks. Aside from the center dense block connecting the two paths, contracting dense blocks are followed by a  $1 \times 1 \times 1$  convolutional layer and a max pooling layer referred to as transition down blocks, and expanding dense blocks are preceded with a  $3 \times 3 \times 3$  transpose convolution with stride 2 known as transition up blocks (Jegou et al., 2017). All convolutional layers in the network are followed by batch normalization and ReLU non-linear layers. Dropout rate of 0.2 is used only after  $3 \times 3 \times 3$  convolutional layers of each dense block. As the final layer a  $1 \times 1 \times 1$  convolution with a sigmoid or softmax output is used, which is discussed later.

### 3.2. Training strategy

While our deep two-channel 3D FC-DenseNet architecture used two extra downsampling and upsampling convolutional layers (purple and red layers in Figure 1) to preserve higher fields of view with more overall number of skip connections than other DenseNet variants (Bui et al., 2017; Dolz et al., 2018), in this work we mainly focused on innovative training strategies to facilitate network training and improve performance. These innovations constitute two training approaches, i.e. single-label multi-class and exclusive multi-label multi-class training, with asymmetric similarity loss functions based on  $F_\beta$  scores (Hashemi et al., 2019), use of large image patches as input, and a patch prediction fusion strategy, which are discussed here.

#### 3.2.1. SINGLE-LABEL MULTI-CLASS

Often in machine learning and deep learning tasks, all labels in a dataset are mutually exclusive which is also the case for the iSeg dataset. This is called a single-label multi-class problem where each voxel can only have one label inside a multi-class environment. One of the most important decisions in a network is the choice of the classification layer. The usual choice for this type of classification for image segmentation is a softmax layer which is a normalized exponential function and a generalization of the logistic function forcing probability values to be in the range of [0,1] with the total class probability sum of 1. Softmax assumes independability of each class to other classes, which is theoretically accurate in the case of iSeg labels (CSF, GM, WM). However, because of human error in generating accurate ground truth labels as well as the isointense specification of GM and WM classes in 6-month infant MRIs, incorporating this theory could result in complications on the border voxels of the two labels where the intensities are most analogous.

In the single-label approach we trained the network the more popular way to learn all labels together with a softmax activation function as shown in Figure 2(a), where the highest probability class was selected for each voxel. Even though we used an asymmetric loss function to account for data imbalance (discussed later), the network applied the required precision-recall asymmetry mostly on labels with higher level of occurrence since all the labels were trained together. In this case the GM label being the most prevalent class (46.7% of all labeled voxels), receives higher recall than the other labels (21.84% CSF and 31.45% WM prevalence). Considering both the level of occurrence as well as the isointense aspect of infant brain MRIs, the WM class would receive the least recall among all labels. Therefore, we aimed to exploit other strategies, in particular exclusive independent probability estimation using a multi-label multi-class strategy to better balance the training.

### 3.2.2. MULTI-LABEL MULTI-CLASS

Unlike single-label problems where voxels can only have one label, in multi-label multi-class problems each voxel has the potential to have multiple labels in a multi-class environment. These types of tasks require prediction of multiple labels per voxel. By using softmax as the activation function, a constant threshold cannot be used practically because the probabilities are not evenly distributed for every patch or image. Therefore, some sort of binary classification or output function is needed; such as the sigmoid function:

$$\sigma_z^{\text{softmax}} = \frac{e^z}{\sum_{k=1}^n e^k} , \quad \sigma_z^{\text{sigmoid}} = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z} = \frac{e^z}{e^0 + e^z} \quad (2)$$

where  $\sigma$  denotes the output of the softmax and sigmoid functions,  $z$  is the output for label  $z$  before activation,  $k$  is the output for each label  $\in [1, n]$  before activation and  $n$  is the total number of labels. Sigmoid is a special case of softmax for only two classes (i.e. 0 and z) which models the probability of classes as Bernoulli distributions and independent from other class probabilities. In the multi-label approach, instead of training all the class labels to a probability sum of 1, we scale each class probability separately between [0,1] so we can use a constant threshold to extract labels. The multi-label multi-class approach has two other advantages: 1) different loss functions and hyper-parameters can be used for distinct training of classes; and 2) calculating sigmoid is less computationally cumbersome for a processing unit compared to softmax especially for large number of labels.

### 3.2.3. EXCLUSIVE MULTI-LABEL MULTI-CLASS

Since we decided to use a less complex cost function and train the class labels independently, there was no reason to train on both of the isointense labels, especially as the classes were mutually exclusive. In fact reducing one of the classes helps the network focus its attention to one label while eliminating the effect of biased learning towards a class with higher prevalence. This way, the model has an easier task of learning subtle differences between nearly indistinguishable classes such as GM and WM in isointense infant brain MRI segmentation. This can potentially be generalizable to every combination of extremely hard to detect, unbalanced, and mutually exclusive class labels, excluding the one with more occurrences and training on the other while reducing the number of network parameters and achieving better performance. To this end, for the iSeg data, as shown in Figure 2(b) we removed GM from the training cycle and trained the CSF and WM classes against non-CSF and non-WM labels using the Sigmoid activation function with differently balanced similarity loss functions discussed in the next section. The GM labels were concluded from the compliment of the already predicted CSF and WM labels.

### 3.2.4. LOSS FUNCTION

To better deal with data imbalance, we used an extension of the idea of using Dice similarity loss (Milletari et al., 2016), based on the  $F_\beta$  scores (Hashemi et al., 2019) defined as:

$$F_\beta = \frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} = \frac{TP}{TP + (\frac{\beta^2}{1+\beta^2})FN + (\frac{1}{1+\beta^2})FP} \quad (3)$$

Where  $TP$ ,  $FN$ , and  $FP$  are true positives, false negatives, and false positives, respectively. While Dice similarity is a harmonic mean of precision and recall,  $F_\beta$  allows balancing between precision

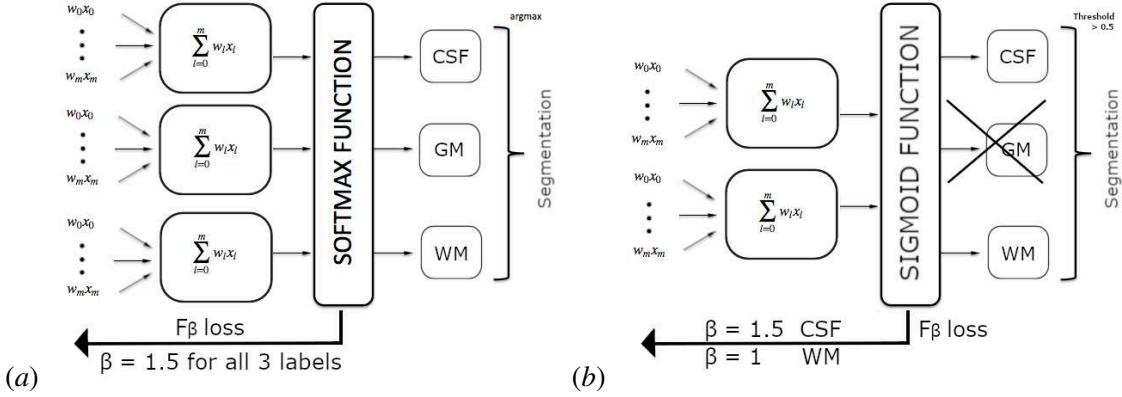


Figure 2: Examples of the a) Single-label approach with the Softmax activation function and single loss function for all labels, and b) Exclusive Multi-label approach with the Sigmoid activation function and multiple loss function configurations for different labels in iSeg segmentation task.

and recall by an appropriate choice of the hyper-parameter  $\beta$ . The selection method for  $\beta$  values based on class prevalence is explained in Appendix A.

### 3.2.5. 3D LARGE PATCHES AND PATCH PREDICTION FUSION

Rather than training on full-size, two-channel (T1- and T2-weighted MRI) input images, we extracted and used large 3D two-channel image patches as inputs and augmented the training data at the level of large patches. This had several advantages including efficient use of memory, intrinsic data augmentation, and the design of an image size-independent model. Previously in the Network Architecture section, we mentioned that large patches of  $128 \times 128 \times 128$  were selected from the image and were immediately downsampled through a convolutional layer within the network to the size of  $64 \times 64 \times 64$  in order to preserve higher receptive field while adapting to GPU memory restrictions. Nonetheless, accuracy near patch borders was relatively low mainly because of the effective receptive field of patches. To circumvent this problem while fusing patch predictions in both training and testing, we exploited a weighted soft voting approach (Hashemi et al., 2019) using second-order spline functions placed at the center of patches. Patches were selected for prediction using 50% overlaps. Each patch was rotated 180 degrees once in each direction for augmentation in both training and testing, and the final result was calculated through the fusion of predictions by all overlapping patches and their augmentations (32 predictions per voxel).

### 3.3. Experimental design

We trained and tested our 3D FC-DenseNet with  $F\beta$  loss layer to segment isointense infant brains. T1- and T2-weighted MRI of 10 subjects were used as input, where we used five-fold cross-validation in training. There was not any pre-processing involved as the images were already skull-stripped and registered. The two MRI images of each subject were normalized through separately dividing each voxel value by the mean of non-zero voxels in each image. This way the whole brain (excluding background) in each modality was normalized to unit mean. Our 3D FC-DenseNet

was trained end-to-end. Cost minimization on 2,500 epochs was performed using ADAM optimizer ([Kingma and Ba, 2014](#)) with an initial learning rate of 0.0005 multiplied by 0.9 every 500 steps. The training time for this network was approximately 14 hours on a workstation with Nvidia Geforce GTX1080 GPU.

Validation and test volumes were segmented using feed-forward through the network. The output of the last convolutional layer with softmax non-linearity consisted of a probability map for CSF, GM and WM tissues. For the sigmoid version of the network, it contained only the CSF and WM tissues. In the softmax approach (single-label multi-class), the class with maximum probability among all classes was selected as the segmentation result for each voxel, while in the sigmoid approach (exclusive multi-label multi-class) voxels with computed probabilities  $\geq 0.5$  were considered to belong to the specific tissue class (CSF or WM) and those with probabilities  $< 0.5$  were considered non tissue. For voxels with both CSF and WM probabilities of  $\geq 0.5$  the class with higher probability was selected. Finally, GM labels were generated based on the compliment of predicted CSF and WM class labels. For evaluation, following iSeg, we report the Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and the Average Surface Distance (ASD) all defined in [Appendix B](#).

## 4. Results

To evaluate the effect of our proposed exclusive multi-label multi-class training strategy compared to the single-label (without exclusive multi-label) method, we trained our FCN with single- and multi-label designs and calculated cross-validation DSC. The characteristics and performance metrics of our two trained models are compared in [Appendix C Table S1](#), along with a comparison of other methods on a different validation set from ([Bui et al., 2017](#)). Paired sample t-test between exclusive multi-label and single-label configurations confirmed that differences were statistically significant ( $p < 0.05$ ) among DSC, sensitivity and specificity results of CSF, GM and WM. The results in the top part of [Table S1](#) and our results in the bottom part should not be directly compared as they are on different validation sets. The actual comparison based on official iSeg test set are reported in [Table 1](#).

The official challenge test results on challenge website, reported in [Table 1](#) for the top performing teams, show that our approach outperformed all first and second round submissions including the DenseNet based methods of DenseSeg ([Bui et al., 2017](#)) and HyperDenseNet ([Dolz et al., 2018](#)). According to the DSC and ASD performance metrics our exclusive multi-label method performed better than all other methods for all (CSF, GM and WM) classes. The results of the HD score, however, were not consistent; nonetheless the HD score is not an appropriate performance measure for segmentation of complex shapes based on the comprehensive discussion and evaluation in ([Taha and Hanbury, 2015](#)) ([Appendix B](#)). Overall, official challenge results show improved segmentation in iSeg using our method, which is attributed to 1) our 3D FC-DenseNet architecture which is deeper than previous DenseNets with more overall skip-layer connections and less number of network parameters; and more importantly 2) our proposed exclusive multi-label training with  $F_\beta$  loss functions that made a better balance between precision and recall in training the network. [Figure 3](#) shows prediction results of a subject from one of the validation folds for our two training methods compared to the ground truth. Visual assessment and the DSC scores on all labels consistently show that the best results were achieved by our exclusive multi-label model.

Table 1: Official iSeg test set results of the top ranking teams. The best values for each metric have been highlighted in bold. Our exclusive multi-label method outperformed the first and second ranked teams (Bui et al., 2017; Dolz et al., 2018) at the time of the challenge and stands, overall, in the first place among all first and second round submissions through December 2018 ([iSeg first round](#), [iSeg second round](#)). Note that the HD metric is not considered a reliable performance metric for medical image segmentation (Taha and Hanbury, 2015) as it is very susceptible to outliers.

Teams (Published)	CSF			GM			WM		
	DSC	HD	ASD	DSC	HD	ASD	DSC	HD	ASD
BCH CRL Imagine (ours)	<b>96.0</b>	<b>8.85</b>	<b>0.11</b>	<b>92.6</b>	9.55	<b>0.31</b>	<b>90.7</b>	7.1	<b>0.36</b>
MSL SKKU (2nd rnd)	95.8	9.11	0.116	92.3	6.0	0.32	90.4	6.62	0.375
MSL SKKU (1st rnd)	95.8	9.07	0.116	91.9	5.98	0.33	90.1	<b>6.44</b>	0.39
LIVIA (2nd rnd)	95.6	9.42	0.12	92.0	<b>5.75</b>	0.33	90.1	6.66	0.38
LIVIA (1st rnd)	95.7	9.03	0.138	91.9	6.42	0.34	89.7	6.98	0.38
Bern IPMI	95.4	9.62	0.127	91.6	6.45	0.34	89.6	6.78	0.4
nic vicorob	95.1	9.18	0.137	91.0	7.65	0.37	88.5	7.15	0.43

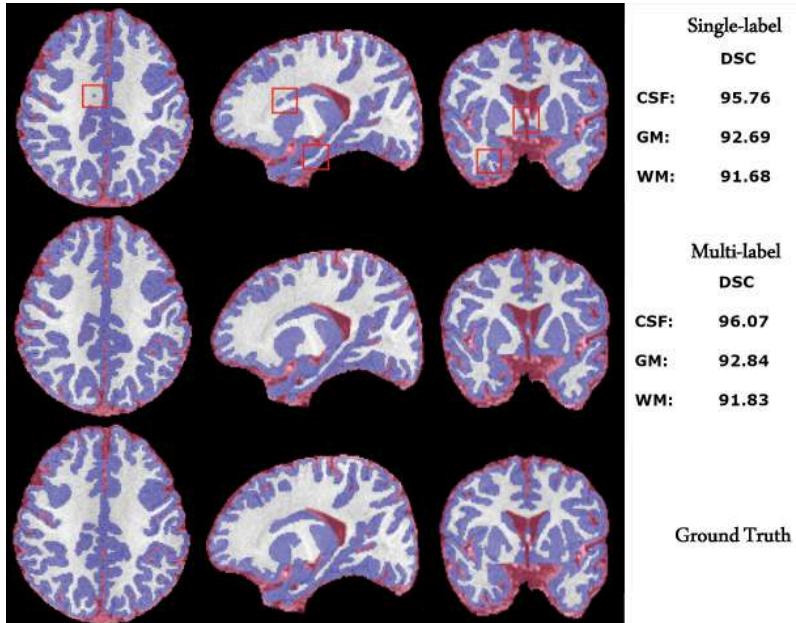


Figure 3: Segmentation results of 3D FC-DensNet of subject 1 in the validation set with Exclusive Multi-label and without Exclusive Multi-label (Single-label). Dice scores for each class label are also shown for each image. Red squares highlight some areas of differences between the two approaches.

## 5. Conclusion

We introduced a new deep densely connected network (Huang et al., 2017) based on (Bui et al., 2017; Jegou et al., 2017), called 3D FC-DenseNet that has more depth, more total number of skip connections and less network parameters than its predecessors, while having the capability of including 8 times the regular patch sizes ( $128 \times 128 \times 128$  vs  $64 \times 64 \times 64$ ) due to its early downsampling and late upsampling layers. To train this deep network we used similarity  $F_\beta$  loss functions that generalized the Dice similarity, and achieved better precision-recall trade-off and thus improved performance in segmentation. We designed two pipelines for training, a single-label (regular network without exclusive multi-label) and an exclusive multi-label procedure. Experimental results in 6-month old infant brain MRI segmentation show that performance evaluation metrics (on the validation set) improved by using exclusive multi-label rather than single-label training. The loss function was designed to weigh recall higher than precision (at  $\beta = 1.5$ ) for CSF, while using equal precision-recall ratio ( $\beta = 1$ ) for WM labels against GM, based on class prevalence in the training set. Official test results based on DSC and ASD scores on the iSeg challenge data show that our method generated the best results in isointense infant brain MRI segmentation, improving the results of all previous DenseNet-based methods (Bui et al., 2017; Dolz et al., 2018).

## 6. Acknowledgements

This study was supported in part by a Technological Innovations in Neuroscience Award from the McKnight Foundation and the National Institutes of Health grants R01EB018988, R01NS079788, and R01NS106030.

## References

- Tom Brosch, Youngjin Yoo, Lisa YW Tang, David KB Li, Anthony Traboulsee, and Roger Tam. Deep convolutional encoder networks for multiple sclerosis lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–11. Springer, 2015.
- Toan Duc Bui, Jitae Shin, and Taesup Moon. 3d densely convolutional networks for volumetric segmentation. *arXiv preprint arXiv:1709.03199*, 2017.
- Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, 170:446–455, 2017.
- Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.
- Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 2018.
- Seyed Raein Hashemi, Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, Sanjay P Prabhu, Simon K Warfield, and Ali Gholipour. Asymmetric loss functions and deep densely-connected networks

for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. *IEEE Access*, 7:1721–1735, 2019.

Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

Simon Jegou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017.

Konstantinos Kamnitsas, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.

Pim Moeskops, Max A Viergever, Adriënne M Mendrik, Linda S de Vries, Manon JNL Benders, and Ivana Išgum. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE transactions on medical imaging*, 35(5):1252–1261, 2016.

Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.

Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Auto-context convolutional neural network (Auto-Net) for brain extraction in magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 36:2319–2330, 2017.

Seyed Sadegh Mohseni Salehi, Seyed Raein Hashemi, Clemente Velasco-Annis, Abdelhakim Ouaalam, Judy A. Estroff, Deniz Erdogmus, Simon K. Warfield, and Ali Gholipour. Real-time automatic fetal brain extraction in fetal mri by deep learning. *IEEE International Symposium on Biomedical Imaging*, pages 720–724, 2018.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, 2015.

Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.

Christian Wachinger, Martin Reuter, and Tassilo Klein. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 170:434–445, 2017.

Li Wang, Feng Shi, Pew-Thian Yap, Weili Lin, John H. Gilmore, and Dinggang Shen. Longitudinally guided level sets for consistent tissue segmentation of neonates. *Human Brain Mapping*, 34(7):956–972, 2013.

Li Wang et al. Benchmark on automatic 6-month-old infant brain segmentation algorithms: The iseg-2017 challenge. *IEEE Transactions on Medical Imaging*, 2019.

Lequan Yu, Jie-Zhi Cheng, Qi Dou, Xin Yang, Hao Chen, Jing Qin, and Pheng-Ann Heng. Automatic 3d cardiovascular mr segmentation with densely-connected volumetric convnets. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 287–295, 2017.

Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji, and Dinggang Shen. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, 108:214–224, 2015.

## Appendix A. Hyper-parameter selection

The values of  $\beta$  for each class are selected automatically in training based on the ratio of the number of instances per every other class over the number of instances for all classes being equal to the coefficient of all false negatives in equation (3):

$$\frac{\beta_z^2}{1 + \beta_z^2} = \frac{\sum_{k=1}^n N_k - N_z}{\sum_{k=1}^n N_k} + \lambda \implies \beta_z = \sqrt{\frac{(1 + \lambda) \sum_{k=1}^n N_k - N_z}{N_z - \lambda \sum_{k=1}^n N_k}} \quad (4)$$

which we saw fit regarding the necessary sensitivity rate for each class based on the complement of its portion on all classes.  $\beta_z$  denotes the chosen value for the  $\beta$  hyper-parameter for label  $z$ ,  $N_z$  corresponds to the total number of labels for class  $z$ ,  $n$  is the number of classes and  $\lambda$  is an extra recall hyper-parameter which we set to 0.1 for this experiment. If we assume  $\lambda$  of 0, then equation (4) becomes the square root of the reverse ratio between the target label and all other labels:

$$\beta_z \Big|_{\lambda=0} = \sqrt{\frac{\sum_{k=1}^n N_k - N_z}{N_z}} \quad (5)$$

Based on prevalence rates of 21.84% for CSF and 31.45% for WM, and  $\lambda = 0.1$ ,  $\beta$  values of 1.5 and 1 were approximated and used in this study for CSF and WM classes, respectively.

## Appendix B. Evaluation metrics

To evaluate and compare the performance of the network against state-of-the-art in isointense infant brain MRI segmentation, we report the Dice Similarity Coefficient (DSC):

$$\text{DSC} = \frac{2 |P \cap R|}{|P| + |R|} = \frac{2TP}{2TP + FP + FN} \quad (6)$$

which is equivalent to the  $F_1$  score calculated by setting  $\beta = 1$  in Equation (3).  $TP$ ,  $FP$ , and  $FN$  are the true positive, false positive, and false negative rates, respectively; and  $P$  and  $R$  are the predicted and ground truth labels, respectively. In the iSeg challenge, in addition to the DSC score, Hausdorff Distance (HD) and Average Surface Distance (ASD):

$$\text{HD} = \max \left\{ \max_{q \in R} d(q, P), \max_{q \in P} d(q, R) \right\}, \quad \text{ASD} = \frac{1}{|R| + |P|} \left( \sum_{q \in R} d(q, P) + \sum_{p \in P} d(p, R) \right) \quad (7)$$

were also reported in the test set results, where  $d(q, P)$  denotes the point-to-set distance:  $d(q, P) = \min_{p \in P} \|q - p\|$ , with  $\|\cdot\|$  presenting the Euclidean distance and  $|\cdot|$  denoting the cardinality of a set.

Average Surface Distance (ASD), also known as Mean Surface Distance (MSD) is the average of all the distances from points on the boundary of  $P$  to the boundary of  $R$  and vice versa, while HD only accounts for the maximum distances between predictions and ground truths. According to (Taha and Hanbury, 2015), HD is generally sensitive to outliers and because noise and outliers are common in medical segmentations, it is not recommended to use HD directly. For example, broken lines that frequently occur in HD calculation on complex shapes increase the HD measure and cause mismatches. Consequently, the use of HD is highly discouraged and is not an appropriate and reliable unit of measure in medical image segmentation where it involves point-to-set matching on complex shapes, which is a procedure that is prone to errors and is susceptible to outliers.

## Appendix C. Tables and Figures

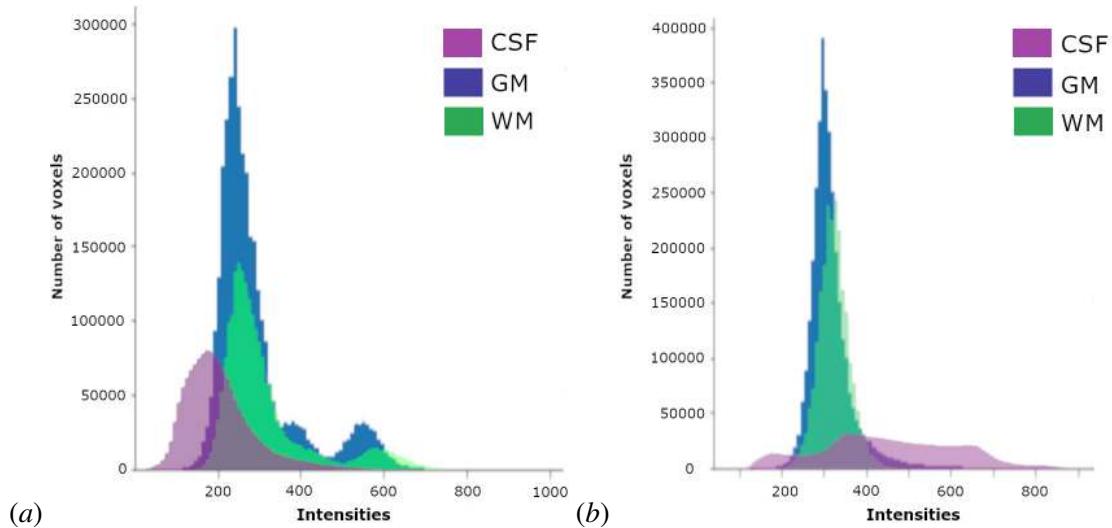


Figure S1: Intensity distributions of all three classes on a) T1-weighted and b) T2-weighted MRI scans on all images in the iSeg training set. At 6 months of age, the intensity values of the white matter (WM) and gray matter (GM) in the infant brain, considered in iSeg, largely overlap. This makes WM-GM segmentation on these images very challenging. Given both scans the cerebrospinal fluid (CSF) shows much better separation from WM and GM based on image intensity values.

Table S1: Average performance metrics (on validation sets) of several state-of-the-art methods trained and evaluated on the iSeg challenge dataset. The best values for each metric have been highlighted in bold. The top three methods in the table are derived from (Bui et al., 2017) with training process and cross validation folds that are different from our methods in the bottom two rows, so the top and bottom parts of the table cannot be directly compared. Comparable results on the official challenge test dataset are shown in Table 1. This table shows relative performance of DenseNets, 3D Unet, and DenseVoxNet style network architectures and their depth and number of parameters. Paired sample t-test between our exclusive multi-label and single-label trained models (bottom two rows) confirmed that differences were statistically significant ( $p < 0.05$ ).

Method	DSC			Depth	Params
	CSF	GM	WM		
3D Unet (Çiçek et al., 2016)	94.44	90.73	89.57	18	19M
DenseVoxNet (Yu et al., 2017)	93.71	88.51	85.46	32	4.34M
DenseSeg - MSK SKKU (Bui et al., 2017)	<b>94.69</b>	<b>91.57</b>	<b>91.25</b>	47	1.55M
3D FC-DenseNet (Single-label)	94.86	91.18	89.27	60	1.5M
3D FC-DenseNet Exclusive Multi-label	<b>95.19</b>	<b>91.79</b>	<b>90.37</b>	60	1.4M

# Dynamic MRI Reconstruction with Motion-Guided Network

**Qiaoying Huang**

QH55@CS.RUTGERS.EDU

**Dong Yang**

DON.YANG.MECH@GMAIL.COM

**Hui Qu**

HUI.QU@CS.RUTGERS.EDU

**Jingru Yi**

JY486@CS.RUTGERS.EDU

**Pengxiang Wu**

PW241@CS.RUTGERS.EDU

**Dimitris Metaxas**

DNM@CS.RUTGERS.EDU

*Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA*

## Abstract

Temporal correlation in dynamic magnetic resonance imaging (MRI), such as cardiac MRI, is informative and important to understand motion mechanisms of body regions. Modeling such information into the MRI reconstruction process produces temporally coherent image sequence and reduces imaging artifacts and blurring. However, existing deep learning based approaches neglect motion information during the reconstruction procedure, while traditional motion-guided methods are hindered by heuristic parameter tuning and long inference time. We propose a novel dynamic MRI reconstruction approach called MODRN that unitizes deep neural networks with motion information to improve reconstruction quality. The central idea is to decompose the motion-guided optimization problem of dynamic MRI reconstruction into three components: dynamic reconstruction, motion estimation and motion compensation. Extensive experiments have demonstrated the effectiveness of our proposed approach compared to other state-of-the-art approaches.

## 1. Introduction

Dynamic magnetic resonance imaging (MRI) is critical in clinical applications, such as cardiovascular and pulmonary. However, high spatiotemporal resolution reconstruction from under-sampled MRI k-space data is still very challenging due to the lack of gold-standard to clinical practice and strong dependence on the parameters tuning (Yang et al., 2016). Many works have been proposed to tackle these challenges. Some of them argued that motion plays a leading role in dynamic reconstruction because correlation and redundancy exist along the temporal dimension, such as cardiac moves relatively periodically against the static background (Jung et al., 2010; Jung and Ye, 2010; Chen et al., 2018). In other computer vision problems, such as video super resolution, motion estimation and compensation are also powerful techniques to further explore the temporal correlations between video frames (Caballero et al., 2017; Makansi et al., 2017).

Traditional Compressed Sensing (CS) approaches have dominated dynamic reconstruction in the past few years. Some studies have successfully incorporated the physical motion into the CS schemes to improve reconstruction performance, either by refining the results after the image reconstruction step (Jung et al., 2009; Bai and Brady, 2011) or embedding the motion estimation into the reconstruction process (Gigengack et al., 2012; Chen et al., 2018). For both methods, displacement motion fields or optical flow are calculated to estimate motion between image pairs. According to the form of motion constraint, many algorithms were proposed, among which, Horn-Schunck (Horn and Schunck, 1981) and Lucas-Kanade (Lucas and Kanade, 1981) are widely used.

However, heuristic parameter tuning and long reconstruction time are major drawbacks of these methods. Additionally, motion estimation itself is still a very difficult task and often produces inaccurate results if object movement is large or fast.

Recent advances in deep learning technique have sparked the new research interests in MRI reconstruction. Deep convolutional neural network was proposed to learn mapping directly from k-space data to fully-sampled reconstructed image, which introduced an interesting way for MRI reconstruction (Zhu et al., 2018). Compared to CS-based methods, deep learning approaches are faster in inference and learn the implicit prior automatically based on the training data (Sun et al., 2016; Schlemper et al., 2017, 2018; Lønning et al., 2018; Qin et al., 2018; Huang et al., 2018). DC-CNN was introduced in (Schlemper et al., 2017), where a differentiable data consistency (DC) layer was added into a deep cascaded convolution neural network for both 2D or dynamic reconstruction. Qin et al. (Qin et al., 2018) efficiently modeled the recurrence of the iterative reconstruction stages by a recurrent network. However, most of them were derived for 2D reconstruction problem. The deep learning based dynamic MRI reconstruction problem is largely unsolved yet critical in clinical scenarios. Especially, all existing state-of-the-art methods did not take motion information into consideration that may generate blurry or temporal inconsistent results.

To tackle the aforementioned problems of both traditional and deep learning methods, in this study, we develop a Motion-guided Dynamic Reconstruction Network (MODRN) that utilizes motion estimation and motion compensation (ME/MC) to improve the reconstruction quality for spatiotemporal imaging. Different from traditional motion estimation algorithms which may fail in low resolution and weak contrast, we utilize the unsupervised deep learning based optical flow estimation (Ren et al., 2017; Meister et al., 2017), which is more robust and accurate in different scenarios. To the best of our knowledge, this is the first work that embeds motion information into deep neural network for dynamic MRI reconstruction. The contribution of this work are three folds. Firstly, we derive a recurrent neural network from the optimization procedures of model-based dynamic reconstruction, which simultaneously links the relationship of data over time and iterations. Secondly, we introduce an unsupervised deep learning based motion estimation method to learn the motion between the reference image and the reconstructed image by using the combination of forward, backward and neighboring loss. Finally, we present a motion compensation component for refining reconstructed image guided by the learned motion.

## 2. Methodology

In this section, we start with the formulation of dynamic MRI reconstruction problem, followed by detailed description of our proposed method called Motion-guided Dynamic Reconstruction Network (MODRN).

### 2.1. Problem Formulation

Given a sequence of under-sampled k-space data  $\{y_t\}_{t \in [T]}$  of  $T$  frames, the dynamic MRI reconstruction problem is to predict reconstructed images  $\{z_t\}_{t \in [T]}$  from  $\{y_t\}$ , which can be formalized as an optimization problem:  $\operatorname{argmin}_{\{z_t\}} \mathcal{L}_{rec}(\{z_t\})$ , where

$$\mathcal{L}_{rec}(\{z_t\}) = \sum_{t=1}^T \frac{1}{2} \|F_u(z_t) - y_t\|_2^2 + \lambda R(z_t). \quad (1)$$

The term  $\|F_u(z_t) - y_t\|_2^2$  is used to guarantee data consistency by restricting the reconstructed image  $z_t$  to be close to the input measurement  $y_t$ .  $F_u(\cdot)$  is an operator that transforms image-domain  $z_t$  into Fourier domain followed by undersampling.  $R(\cdot)$  is a regularization function that depends on the prior knowledge of the input  $\{y_t\}$ . Common choices include sparsity in transformed domain (Lingala et al., 2011), total variation (TV) penalties (Knoll et al., 2012) and low-rank property (Trzasko et al., 2011).  $\lambda$  is a weighting factor.

In order to capture anatomical motion in the dynamic MRI acquisition, it is natural to incorporate motion estimation/motion compensation (ME/MC) technique in the reconstruction process (Jung et al., 2010). Specifically, based on the brightness constancy assumption (Horn and Schunck, 1981), for a temporal 2D image  $z_t(x, y, t)$  with small movement  $(\Delta x, \Delta y, \Delta t)$  with respect to the next frame, we add the following motion estimation constraint to the objective function (1):

$$\mathcal{L}_{me}(\{v_t\}) = \sum_{t=1}^{T-1} \left\| \nabla z_t^\top v_t + \frac{\partial z_t}{\partial t} \right\|_1 + \delta \|v_t\|_1, \quad (2)$$

where  $\nabla z_t(x, y) = \left( \frac{\partial z_t}{\partial x}, \frac{\partial z_t}{\partial y} \right)$  are the derivatives of image  $z_t$  at position  $(x, y)$ , and  $v_t(x, y) = \left( \frac{\Delta x}{\Delta t}, \frac{\Delta y}{\Delta t} \right)$  is the estimated displacement motion fields or optical flow.

Furthermore, given the estimated motion field  $v_t$ , the reconstructed image  $z_t$  can be refined through MC process, *i.e.*  $c_t = \text{MC}(z_t, z_1, z_T) + r_t$ , where  $c_t$  is the motion-compensated reconstructed image and  $r_t$  is a residual term for better exploiting temporal redundancy. Therefore, we can derive the motion compensation constraint as follows.

$$\mathcal{L}_{mc}(\{r_t\}) = \sum_{t=1}^{T-1} \frac{1}{2} \|F_u(c_t) - y_t\|_2^2. \quad (3)$$

By combining with two motion-based constraints of Equations (2) and (3), the motion-guided dynamic MRI reconstruction problem is defined as:

$$\underset{\{z_t, v_t, r_t\}}{\operatorname{argmin}} \mathcal{L}_{rec}(\{z_t\}) + \eta \mathcal{L}_{me}(\{v_t\}) + \zeta \mathcal{L}_{mc}(\{r_t\}). \quad (4)$$

The solution to problem (4) is non-trivial and traditional CS-based algorithms are usually computationally expensive and require long running time for hyper-parameter tuning. Recent advances in deep learning provide an alternative way for efficient MRI reconstruction, but very few works focused on the dynamic reconstruction problem and they only targeted for the simpler problem (1) without considering motion information. To this end, we propose a deep learning based method called Motion-guided Dynamic Reconstruction Network (MODRN) to solve the problem (4).

## 2.2. Motion-guided Dynamic Reconstruction Network

Our method dissects the motion-guided dynamic reconstruction problem into three closely-connected parts: (i) Dynamic Reconstruction (DR) component for estimating initial reconstructed image from Equation (1), (ii) Motion Estimation (ME) component for generating motion information through Equation (2), and (iii) Motion Compensation (MC) component for refining reconstructed image guided by learned motion based on Equation (3).

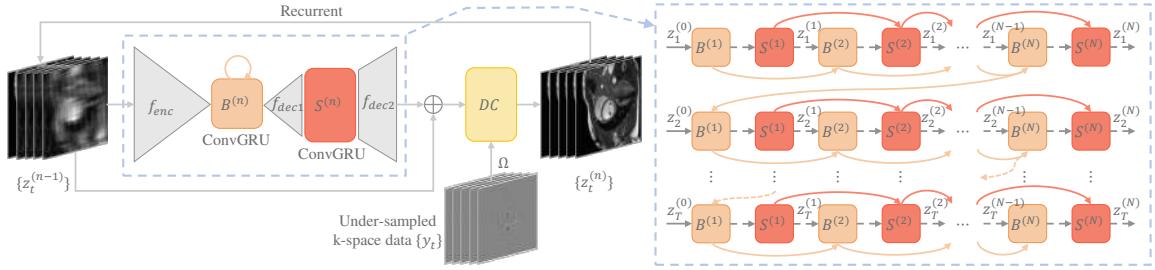


Figure 1: Architecture of DR component and the workflow of ConvGRU units  $B$  and  $S$ .

### 2.2.1. DYNAMIC RECONSTRUCTION

Instead of directly solving Equation (1), we adopt an iterative process through DR component to approximate reconstructed images  $z_t$ . Formally, given under-sampled k-space data  $\{y_t\}_{t \in [T]}$  with sampled mask  $\Omega$ , DR component learns to reconstruct images in  $N$  iterations:

$$z_t^{(n)} = DR(z_t^{(n-1)}, y_t, \Omega) \Leftrightarrow \begin{cases} x_{bt}^{(n)}, b_t^{(n)} = B(f_{enc}(z_t^{(n-1)}), b_t^{(n-1)}) \\ x_{st}^{(n)}, s_t^{(n)} = S(f_{dec1}(x_{bt}^{(n)}), s_t^{(n-1)}) \\ z_t^{(n)} = DC(f_{dec2}(x_{st}^{(n)}) + z_t^{(n-1)}, y_t, \Omega) \end{cases}, n \in [N]. \quad (5)$$

where  $z_t^{(0)}$  is zero-filling image and  $z_t^{(n)}$  is the reconstructed image of  $y_t$  after iteration  $n$ .  $B$  and  $S$  are two ConvGRU (Ballas et al., 2015) units that respectively output features  $x_{bt}^{(n)}$  and  $x_{st}^{(n)}$  together with hidden states  $b_t^{(n)}$  and  $s_t^{(n)}$ .  $f_{enc}$  and  $f_{dec1}, f_{dec2}$  are convolutional encoder and decoders in the U-Net (Ronneberger et al., 2015), which is used as the backbone of the DR component to capture coarse-to-fine features of reconstructed images. Equation (5) is visualized in figure 1 for better understanding. One benefit here is that regularization function  $R(\cdot)$  in Equation (1) is now built upon the convolutional network for automated feature learning and hence avoid the requirements of prior knowledge on the selection of  $R$ .  $DC(\cdot)$  is the differentiable DC layer (Schlemper et al., 2017) that takes the same effect as the data consistency term  $\|F_u(z_t) - y_t\|_2^2$  in Equation (1) to force the reconstructed image to be consistent with the input data. It fills the reconstructed image  $z_t$  with the original values of input data  $y_t$  in the Fourier domain by the sampled mask  $\Omega$ .

More importantly, in order to capture dynamic information of image sequence during each iteration, we introduce two kinds of ConvGRU units, *i.e.*  $B$  and  $S$ , inspired by the work (Qin et al., 2018) in Equation (5). The difference between  $B$  and  $S$  is that GRU unit  $S$  is used to improve the performance of image  $z_t$  over  $N$  iterations while the role of  $B$  is to connect dynamic information of neighboring images  $z_{t-1}$  and  $z_t$ , which is implemented by initializing hidden state  $b_t^{(0)}$  as  $b_{t-1}^{(N)}$ . Finally, we impose  $l_1$  loss on the reconstructed images  $\{z_t^N\}$  with respect to ground truth for penalizing.

### 2.2.2. MOTION ESTIMATION

In analogy to Equation (2), the Motion Estimation (ME) component takes as input the sequence of reconstructed images  $\{z_t\}_{t \in [T]}$  and learn to predict displacement motion fields  $\{v_t\}_{t \in [T]}$ . As shown in figure 2, our proposed ME component embraces two parts. One is a FlowNet backboned by

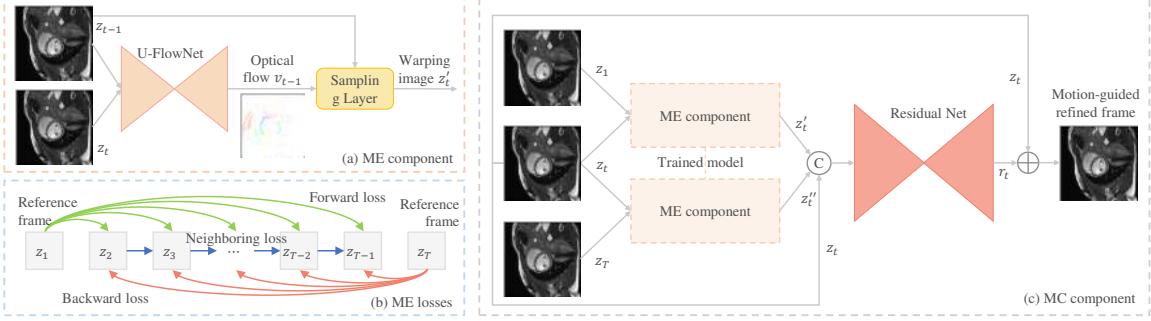


Figure 2: The network architecture of ME/MC components and ME losses.

convolutional U-Net (U-FlowNet) for motion field estimation. The other is a differentiable sampling layer based on Spatial Transformer Network (STN) (Jaderberg et al., 2015), which endows convolutional network with the ability to warp the spatial deformation between images. Unlike traditional optimization algorithms for motion estimation that depend on a strong assumption that the brightness of two frames should be consistent and the movement of the foreground object is small, our method does not succumb to any assumption and hence is more applicable in practical dynamic MRI reconstruction. The performance of ME is heavily affected by noisy input, therefore it is pre-trained with two fully sampled images  $z_{t-1}$  and  $z_t$ . The image pair is first fed to the U-FlowNet, which produces two-channel displacement  $v_{t-1}$  along the  $x$  and  $y$  directions. Then, the sampling layer warps  $z_{t-1}$  towards  $z_t$  by using  $v_{t-1}$  and yields a warping image denoted by  $z'_t$  through differentiable bilinear interpolation. This leads to a natural re-formalization of motion estimation (ME) loss  $\ell_{me}$  between  $z_{t-1}$  and  $z_t$  from Equation (2):

$$\ell_{me}(z_{t-1}, z_t) = \|z'_t - z_t\|_1 + \beta \|v_{t-1}\|_1 + \gamma \|v_{t-1}\|_{TV}. \quad (6)$$

The first term is an image reconstruction loss used to keep the majority of high-frequency parts on images. Two additional regularization terms reinforce constraints on the motion field  $v_{t-1}$ , where  $l_1$  regularization is to suppress unreal large magnitude of displacement and total-variation (TV) regularization is to make the displacement locally smooth.

In addition, the above loss only enforces temporal consistency between consecutive frames, but there is no guarantee for long-term coherence. Therefore, we consider to train the U-FlowNet with three sets of ME losses to capture long-term motion information, as illustrated in figure 2.

$$\mathcal{L}_{me}(\{z_t\}) = \sum_{t=2}^{T-1} \ell_{me}(z_1, z_t) + \sum_{t=2}^{T-1} \ell_{me}(z_t, z_T) + \sum_{t=2}^{T-2} \ell_{me}(z_t, z_{t+1}), \quad (7)$$

where three terms on the right-hand-side are respectively forward ME loss, backward ME loss and neighboring ME loss.

### 2.2.3. MOTION COMPENSATION

Motion Compensated (MC) component is used to refine reconstructed images  $\{z_t\}_{t \in [T]}$  through motion information and to generate motion compensated image  $\{c_t\}_{t \in [T]}$ . By following the work (Jung et al., 2010), during the MC stage, we also add two additional fully sampled reference frames

to learn more accurate displacement motion fields. The pre-trained U-FlowNet is fixed and directly used as an operator in the MC component. As shown in figure 2, the MC component takes a reconstructed image  $z_t$  from the DR component and two reference frame  $z_1$  and  $z_T$  as input. It first retrieves two warping images  $z'_t$  and  $z''_t$  from the ME component by feeding  $z_1, z_t$  and  $z_t, z_T$  respectively. These two images represent forward and backward motion information, which is then concatenated and fed to a residual network to generate residual information  $r_t$ , as described in Equation (3). Finally, the reconstructed image  $z_t$  together with the residual  $r_t$  are summed up to generate the motion-guided refined image  $c_t$ , which is penalized by  $l_1$  loss with respect to the ground truth image.

### 3. Experiment

**Evaluation Dataset:** We experiment with a short-axis (SAX) cardiac dataset composed of 15 patients. Each subject contains around 12 SAX planes and each plane includes 24 phases (2D images) that form a whole cardiac cycle. The image resolution is normalized to  $1.25\text{mm}$  and image size is cropped to  $152 \times 152$  pixels. In order to simulate k-space data, we adopt the same Cartesian under-sampling method as introduced in (Jung et al., 2007), which assumes that sampled mask  $\Omega$  follows a zero-mean Gaussian distribution and keeps 8 center spatial frequencies. We consider two different settings on the dataset respectively with under-sampling rates of 20% (or acceleration rate  $5\times$ ) and 12.5% ( $8\times$ ). For convenience, we refer to these two cases as *Rate 5×* and *Rate 8×*. We perform 3-fold cross-validation in the following experiments that each fold contains 10 training subjects and 5 test subjects.

**Implementation Details:** We implement all the deep learning models with PyTorch and train them on NVIDIA K80. All models are trained for total 80 epochs using Adam optimizer, with initialized learning rate of  $5 \times 10^{-4}$  and decreasing rate of 0.5 for every 20 epochs. Due to hardware limitation, the number of iterations is set to be  $N = 3$  and the length of image sequence is  $T = 12$ .

#### 3.1. Comparison to State-of-the-Art

In this experiment, we evaluate the dynamic reconstruction performance of our proposed methods quantitatively and qualitatively in both cases of *Rate 5×* and *Rate 8×*. We consider three variants of our models: DRN w/o GRU (the one without GRU hidden unit), DRN (the one with DR component

Table 1: Average performance of dynamic MRI reconstruction on the test subjects in both cases of *Rate 5×* and *Rate 8×*. The best results are highlighted in bold font.

Method	NRMSE↓	PSNR↑	SSIM↑	NRMSE↓	PSNR↑	SSIM↑
	5×			8×		
k-t SLR	0.0934	21.0858	0.6794	0.1054	19.9504	0.6193
k-t FOCUSS	0.0766	22.7471	0.6581	0.0879	21.4063	0.5920
k-t FOCUSS+ME/MC	0.0758	22.8139	0.6701	0.0854	21.6547	0.6131
DC-CNN(3D)	0.0360	29.1292	0.8449	0.0513	25.9709	0.7441
DRN w/o GRU	0.0381	28.7187	0.8286	0.0519	25.9120	0.7448
DRN	0.0349	29.5394	0.8502	0.0485	26.5275	0.7687
MODRN	<b>0.0274</b>	<b>32.0403</b>	<b>0.9104</b>	<b>0.0364</b>	<b>29.4774</b>	<b>0.8702</b>

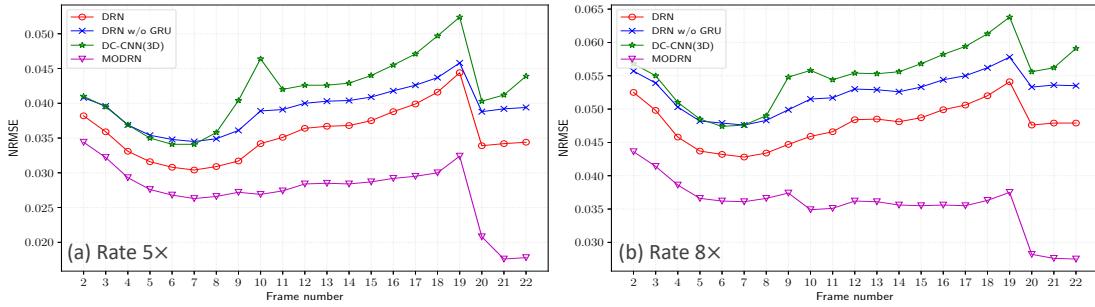


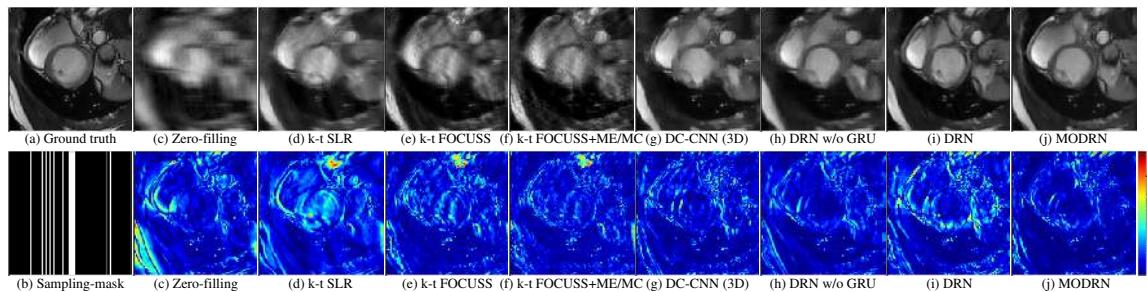
Figure 3: NRMSE curves of deep learning methods within one cardiac cycle in two cases.

only) and MODRN (the complete version). We compare with four state-of-the-art approaches including k-t SLR (Lingala et al., 2011), k-t FOCUSS (Jung et al., 2009), k-t FOCUSS+ME/MC (Jung et al., 2010) and DC-CNN (3D) (Schlemper et al., 2017). The first three are traditional CS-based methods and only k-t FOCUSS+ME/MC includes ME/MC procedures. The last one is also a deep learning based method that explores spatio-temporal information using 3D convolution. Three common quantitative metrics are used: root square mean error (NRMSE), peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM).

**Quantitative Results:** The results of all methods are reported in Table 1. We observe that all our methods consistently outperform four state-of-the-art approaches in both *Rate 5×* and *Rate 8×* cases. In particular, MODRN achieves the best performance for all metrics, mainly attributing to the motion information exploited by ME/MC components. We also find that DRN outperforms DRN w/o GRU by a large margin, which indicates the importance of utilizing dynamic sequence of image.

To further investigate the performance of four deep learning methods, we plot NRMSE values within a complete cardiac cycle of one example in Figure 3. It shows that our method MODRN consistently achieves the smallest error of dynamic reconstruction for the sequence of images. In contrast, the models without ME/MC are unstable along the temporal dimension, especially in the case of DC-CNN(3D). For example, in the case of *Rate 8×*, the gap between DRN and MODRN model become larger, which implies the significance of using motion information.

**Qualitative Results:** We visualize the reconstructed images and error with respect to ground truth of all methods in Figure 4. It is obvious that all CS-based methods have streaking artifacts and larger

Figure 4: Visualization of reconstructed images and errors in the case of *Rate 8×*.

Method	Dice $\uparrow$	HD $\downarrow$
Reference	0.8130	1.9254
Lucas-Kanade	0.8125	1.9577
U-FlowNet-A	0.8297	1.8755
U-FlowNet-B	<b>0.8306</b>	<b>1.8584</b>

Table 2: Motion estimation results.

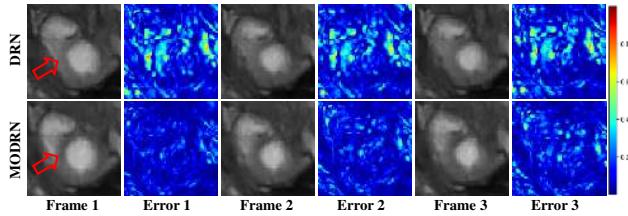


Figure 5: Motion compensation results.

reconstruction error while our MODRN model eliminates the most blurring artifacts and recovers more high-frequency details.

### 3.2. Motion Estimation/Compensation Analysis

First, we consider to evaluate the motion estimation results generated by the U-FlowNet from ME component. Two baseline methods, Reference and Lucas-Kanade, are compared with our U-FlowNet-A (trained with only neighboring loss) and U-FlowNet-B (trained with combined loss). Reference method directly calculates metrics using the segmentation of the target phase and the reference phase. Since it is impractical to obtain the ground truth of optical flow from cardiac MRI, we compute the overlapped area of the myocardium between the targeting image and the warping image. In particular, we calculate the average Dice’s score and Hausdorff Distance between  $z_1$  and other frames,  $z_T$  and other frames and also neighboring frames. The results of 3-fold cross-validation are reported in Table 2. We observe that U-FlowNet-B method achieves the best performance, which indicates that compared with neighboring loss, our combined loss contributes more to accurate motion estimation with large movement between frames.

Second, we compare the quality of motion-guided refined image by MC component of MODRN with that of reconstructed image by DRN alone. The results of three consecutive frames are visualized in Figure 5. We can observe clear improvements of MODRN that its reconstruction error is reduced around cardiac region and no noticeable artifact is generated.

## 4. Conclusion

We present a novel deep learning based approach called MODRN for motion-guided dynamic MRI reconstruction problem. It is featured by a dynamic reconstruction (DR) component for preliminary image reconstruction from under-sampled k-space data, and a motion estimation (ME) component to predict motion of image sequence, which is further exploited by a motion compensation (MC) component to refine the motion-guided reconstructed images. We extensively evaluate our approach on a short-axis cardiac dataset in two settings. The experimental results show the effectiveness of MODRN compared to state-of-the-art methods and prove the significance of motion information from ME/MC components.

## References

- Wenjia Bai and Michael Brady. Motion correction and attenuation correction for respiratory gated pet images. *IEEE transactions on medical imaging*, 30(2):351–365, 2011.

Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015.

Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4778–4787, 2017.

Chong Chen, Barbara Gris, and Ozan Öktem. A new variational model for joint image reconstruction and motion estimation in spatiotemporal imaging. *arXiv preprint arXiv:1812.03446*, 2018.

Fabian Gigengack, Lars Ruthotto, Martin Burger, Carsten H Wolters, Xiaoyi Jiang, and Klaus P Schafers. Motion correction in dual gated cardiac pet using mass-preserving image registration. *IEEE transactions on medical imaging*, 31(3):698–712, 2012.

Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3): 185–203, 1981.

Qiaoying Huang, Dong Yang, Pengxiang Wu, Hui Qu, Jingru Yi, and Dimitris Metaxas. Mri reconstruction via cascaded channel-wise attention network. *arXiv preprint arXiv:1810.08229*, 2018.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

Hong Jung and Jong Chul Ye. Motion estimated and compensated compressed sensing dynamic magnetic resonance imaging: What we can learn from video compression techniques. *International Journal of Imaging Systems and Technology*, 20(2):81–98, 2010.

Hong Jung, Jong Chul Ye, and Eung Yeop Kim. Improved k-t blast and k-t sense using focuss. *Physics in Medicine & Biology*, 52(11):3201, 2007.

Hong Jung, Kyunghyun Sung, Krishna S Nayak, Eung Yeop Kim, and Jong Chul Ye. k-t focuss: a general compressed sensing framework for high resolution dynamic mri. *Magnetic resonance in medicine*, 61(1):103–116, 2009.

Hong Jung, Jaeseok Park, Jaeheung Yoo, and Jong Chul Ye. Radial k-t focuss for high-resolution cardiac cine mri. *Magnetic Resonance in Medicine*, 63(1):68–78, 2010.

Florian Knoll, Christian Clason, Kristian Bredies, Martin Uecker, and Rudolf Stollberger. Parallel imaging with nonlinear reconstruction using variational penalties. *Magnetic resonance in medicine*, 67(1):34–41, 2012.

Sajan Goud Lingala, Yue Hu, Edward DiBella, and Mathews Jacob. Accelerated dynamic mri exploiting sparsity and low-rank structure: kt slr. *IEEE transactions on medical imaging*, 30(5): 1042–1054, 2011.

Kai Lønning, Patrick Putzky, Matthan WA Caan, and Max Welling. Recurrent inference machines for accelerated MRI reconstruction. In *International Conference on Medical Imaging with Deep Learning (MIDL 2018)*, 2018.

Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc. URL <http://dl.acm.org/citation.cfm?id=1623264.1623280>.

Osama Makansi, Eddy Ilg, and Thomas Brox. End-to-end learning of video super-resolution with motion compensation. In *German conference on pattern recognition*, pages 203–214. Springer, 2017.

Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. *arXiv preprint arXiv:1711.07837*, 2017.

Chen Qin, Joseph V Hajnal, Daniel Rueckert, Jo Schlemper, Jose Caballero, and Anthony N Price. Convolutional recurrent neural networks for dynamic mr image reconstruction. *IEEE transactions on medical imaging*, 2018.

Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *AAAI*, volume 3, page 7, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for mr image reconstruction. In *International Conference on Information Processing in Medical Imaging*, pages 647–658. Springer, 2017.

Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503, 2018.

Jian Sun, Huibin Li, Zongben Xu, et al. Deep admm-net for compressive sensing mri. In *Advances in Neural Information Processing Systems*, pages 10–18, 2016.

J Trzasko, A Manduca, and E Borisch. Local versus global low-rank promotion in dynamic mri series reconstruction. In *Proc. Int. Symp. Magn. Reson. Med*, page 4371, 2011.

Alice Chieh-Yu Yang, Madison Kretzler, Sonja Sudarski, Vikas Gulani, and Nicole Seiberlich. Sparse reconstruction techniques in mri: methods, applications, and challenges to clinical adoption. *Investigative radiology*, 51(6):349, 2016.

Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487, 2018.

# Boundary loss for highly unbalanced segmentation

**Hoel Kervadec**\*<sup>1</sup>

**Jihene Bouchtiba**\*<sup>1</sup>

**Christian Desrosiers**<sup>1</sup>

**Eric Granger**<sup>1</sup>

**Jose Dolz**<sup>1</sup>

**Ismail Ben Ayed**<sup>1</sup>

<sup>1</sup> ÉTS Montreal

HOEL.KERVADEC.1@ESTMTL.NET

JIHENE.BOUCHTIBA.1@ENS.ETSMTL.CA

CHRISTIAN.DESROSIERS@ETSMTL.CA

ERIC.GRANGER@ETSMTL.CA

JOSE.DOLZ@ETSMTL.CA

ISMAIL.BENAYED@ETSMTL.CA

## Abstract

Widely used loss functions for convolutional neural network (CNN) segmentation, e.g., Dice or cross-entropy, are based on integrals (summations) over the segmentation regions. Unfortunately, for highly unbalanced segmentations, such regional losses have values that differ considerably – typically of several orders of magnitude – across segmentation classes, which may affect training performance and stability. We propose a *boundary* loss, which takes the form of a distance metric on the space of contours (or shapes), not regions. This can mitigate the difficulties of regional losses in the context of highly unbalanced segmentation problems because it uses integrals over the boundary (interface) between regions instead of unbalanced integrals over regions. Furthermore, a boundary loss provides information that is complimentary to regional losses. Unfortunately, it is not straightforward to represent the boundary points corresponding to the regional softmax outputs of a CNN. Our boundary loss is inspired by discrete (graph-based) optimization techniques for computing gradient flows of curve evolution. Following an integral approach for computing boundary variations, we express a non-symmetric  $L_2$  distance on the space of shapes as a regional integral, which avoids completely local differential computations involving contour points. This yields a boundary loss expressed with the regional softmax probability outputs of the network, which can be easily combined with standard regional losses and implemented with any existing deep network architecture for N-D segmentation. We report comprehensive evaluations on two benchmark datasets corresponding to difficult, highly unbalanced problems: the ischemic stroke lesion (ISLES) and white matter hyperintensities (WMH). Used in conjunction with the region-based generalized Dice loss (GDL), our boundary loss improves performance significantly compared to GDL alone, reaching up to 8% improvement in Dice score and 10% improvement in Hausdorff score. It also yielded a more stable learning process. Our code is publicly available<sup>1</sup>.

**Keywords:** Boundary loss, unbalanced data, semantic segmentation, deep learning, CNN

## 1. Introduction

Recent years have witnessed a substantial growth in the number of deep learning methods for medical image segmentation (Litjens et al., 2017; Shen et al., 2017; Dolz et al., 2018; Ker et al., 2018). Widely used loss functions for segmentation, e.g., Dice or cross-entropy, are based on *regional* integrals, which are convenient for training deep neural networks. In practice, these regional integrals are summations over the segmentation regions of differentiable functions, each invoking directly

---

\* Contributed equally

1. <https://github.com/LIVIAETS/surface-loss>

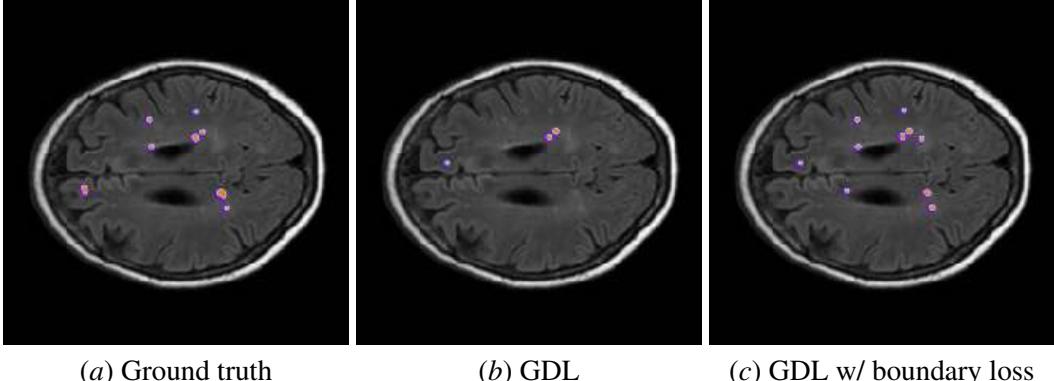


Figure 1: A visual comparison that shows the positive effect of our boundary loss on a validation data from the WMH dataset. Our boundary loss helped recovering small regions that were otherwise missed by the generalized Dice loss (GDL). Best viewed in colors.

the softmax probability outputs of the network. Therefore, standard stochastic optimizers such SGD are directly applicable. Unfortunately, difficulty occurs for highly unbalanced segmentations, for instance, when the size of target foreground region is several orders of magnitude less than the background size. For example, in the characterization of white matter hyperintensities (WMH) of presumed vascular origin, the foreground composed of WMH regions may be 500 times smaller than the background (see the typical example in Fig. 1). In such cases, quite common in medical image analysis, standard regional losses contain foreground and background terms that have substantial differences in their values, typically of several orders of magnitude, which might affect performance and training stability (Milletari et al., 2016; Sudre et al., 2017).

Segmentation approaches based on convolutional neural networks (CNN) are typically trained by minimizing the cross-entropy (CE), which measures an affinity between the regions defined by probability softmax outputs of the network and the corresponding ground-truth regions. The standard regional CE has well-known drawbacks in the context of highly unbalanced problems. It assumes identical importance distribution of all the samples and classes. To achieve good generalization, it requires a large training set with balanced classes. For unbalanced data, CE typically results in unstable training and leads to decision boundaries biased towards the majority classes. Class-imbalanced learning aims at mitigating learning bias by promoting the importance of infrequent labels. In medical image segmentation, a common strategy is to re-balance class prior distributions by down-sampling frequent labels (Havaei et al., 2017; Valverde et al., 2017). Nevertheless, this strategy limits the information of the images used for training. Another common practice is to assign weights to the different classes, inversely proportional to the frequency of the corresponding labels (Brosch et al., 2015; Ronneberger et al., 2015; Kamnitsas et al., 2017; Long et al., 2015; Yu et al., 2017). In this scenario, the standard cross-entropy (CE) loss is modified so as to assign more importance to the rare labels. Although effective for some unbalanced problems, such weighting methods may undergo serious difficulties when dealing with highly unbalanced datasets, as seen with WMH segmentation. The CE gradient computed over the few pixels of infrequent labels is typically noisy, and amplifying this noise with a high class weight may lead to instability.

The well-known Dice overlap coefficient was also adopted as a regional loss function, typically outperforming CE in unbalanced medical image segmentation problems (Milletari et al., 2016, 2017; Wong et al., 2018). Sudre et al. (Sudre et al., 2017) generalized the Dice loss (Milletari et al., 2016) by weighting according to the inverse of class-label frequency. Despite these improvements over CE (Milletari et al., 2016; Sudre et al., 2017), regional Dice losses may undergo difficulties when dealing with very small structures. In such highly unbalanced scenarios, mis-classified pixels may lead to large decreases of the loss, resulting in unstable optimization. Furthermore, Dice corresponds to the harmonic mean between precision and recall, implicitly using the arithmetic mean of false positives and false negatives. False positives and false negatives are, therefore, equally important when the true positives remain the same, making this loss mainly appropriate when both types of errors are equally bad. The recent research in (Salehi et al., 2017; Abraham and Khan, 2018) investigated losses based on the Tversky similarity index in order to provide a better trade-off between precision and recall. It introduced two parameters that control the importance of false positives and false negatives. Other recent advances in class-imbalanced learning for computer vision problems have been adopted in medical image segmentation. For example, inspired by the concept of focal loss (Lin et al., 2018), Dice and Tversky losses have been extended to integrate a focal term, which is parameterized by a value that controls the importance between easy and hard training samples (Abraham and Khan, 2018; Wong et al., 2018). The main objective of these losses is to balance the classes not only by their relative class sizes, but also by the level of segmentation difficulty.

## Contributions

All the above-mentioned losses are *region-based*. In this paper, we propose a *boundary* loss that takes the form of a distance metric on the space of contours (or shapes), not regions. We argue that a boundary loss can mitigate the issues related to regional losses in highly unbalanced segmentation problems. Rather than using unbalanced integrals over the regions, a boundary loss uses integrals over the boundary (interface) between the regions. Furthermore, it provides information that is complimentary to regional losses. It is, however, challenging to represent the boundary points corresponding to the regional softmax outputs of a CNN. This difficulty may explain why boundary losses have been avoided in the context of deep segmentation networks. Our boundary loss is inspired by techniques in discrete (graph-based) optimization for computing gradient flows of curve evolution (Boykov et al., 2006). Following an integral approach for computing boundary variations, we express a non-symmetric  $L_2$  distance on the space of shapes (or contours) as a regional integral, which avoids completely local differential computations involving contour points. This yields a boundary loss expressed as the sum of linear functions of the regional softmax probability outputs of the network. Therefore, it can be easily combined with standard regional losses and implemented with any existing deep network architecture for N-D segmentation.

We evaluated our boundary loss in conjunction with the region-based generalized Dice loss (GDL) (Sudre et al., 2017) on two challenging and highly unbalanced segmentation problems – the Ischemic Stroke Lesion (ISLES) and the White Matter Hyperintensities (WMH) benchmark datasets. The results indicate that the proposed boundary loss provides a more stable learning process, and significantly outperforms GDL alone, yielding up to 8% improvement in Dice score and 10% improvement in Hausdorff score.

## 2. Formulation

Let  $I : \Omega \subset \mathbb{R}^{2,3} \rightarrow \mathbb{R}$  denotes a training image with spatial domain  $\Omega$ , and  $g : \Omega \rightarrow \{0,1\}$  a binary ground-truth segmentation of the image:  $g(p) = 1$  if pixel/voxel  $p$  belongs to the target region  $G \subset \Omega$  (foreground region) and 0 otherwise, i.e.,  $p \in \Omega \setminus G$  (background region)<sup>2</sup>. Let  $s_\theta : \Omega \rightarrow [0, 1]$  denotes the softmax probability output of a deep segmentation network, and  $S_\theta \subset \Omega$  the corresponding segmentation region:  $S_\theta = \{p \in \Omega \mid s_\theta(p) \geq \delta\}$  for some threshold  $\delta$ . Widely used segmentation loss functions involve a *regional integral* for each segmentation region in  $\Omega$ , which measures some similarity (or overlap) between the region defined by the probability outputs of the network and the corresponding ground-truth. In the two-region case, we have an integral of the general form  $\int_{\Omega} g(p)f(s_\theta(p))dp$  for the foreground, and of the form  $\int_{\Omega}(1-g(p))f(1-s_\theta(p))dp$  for the background. For instance, the standard two-region cross-entropy loss corresponds to a summation of these two terms for  $f = -\log(\cdot)$ . Similarly, the generalized Dice loss (GDL) (Sudre et al., 2017) involves regional integrals with  $f = 1$ , subject to some normalization, and is given as follows for the two-region case:

$$\mathcal{L}_{GD}(\theta) = 1 - 2 \frac{w_G \int_{p \in \Omega} g(p)s_\theta(p)dp + w_B \int_{p \in \Omega}(1-g(p))(1-s_\theta(p))dp}{w_G \int_{\Omega}[s_\theta(p) + g(p)]dp + w_B \int_{\Omega}[2 - s_\theta(p) - g(p)]dp} \quad (1)$$

where coefficients  $w_G = 1 / \left( \int_{p \in \Omega} g(p)dp \right)^2$  and  $w_B = 1 / \left( \int_{\Omega}(1-g(p))dp \right)^2$  are introduced to reduce the well-known correlation between the Dice overlap and region size.

Regional integrals are widely used because they are convenient for training deep segmentation networks. In practice, these regional integrals are summations of differentiable functions, each invoking directly the softmax probability outputs of the network,  $s_\theta(p)$ . Therefore, standard stochastic optimizers such SGD are directly applicable. Unfortunately, extremely unbalanced segmentations are quite common in medical image analysis, where, e.g., the size of the target foreground region is several orders of magnitude smaller than the background size. This represents challenging cases because the foreground and background terms have substantial differences in their values, which affects segmentation performance and training stability (Milletari et al., 2016; Sudre et al., 2017).

Our purpose is to build a boundary loss  $\text{Dist}(\partial G, \partial S_\theta)$ , which takes the form of a distance metric on the space of contours (or region boundaries) in  $\Omega$ , with  $\partial G$  denoting a representation of the boundary of ground-truth region  $G$  (e.g., the set of points of  $G$ , which have a spatial neighbor in background  $\Omega \setminus G$ ) and  $\partial S_\theta$  denoting the boundary of the segmentation region defined by the network output. On the one hand, a boundary loss should be able to mitigate the above-mentioned difficulties for unbalanced segmentations: rather than using unbalanced integrals within the regions, it uses integrals over the boundary (interface) between the regions. Furthermore, a boundary loss provides information that is different from and, therefore, complimentary to regional losses. On the other hand, it is not clear how to represent boundary points on  $\partial S_\theta$  as a differentiable function of regional network outputs  $s_\theta$ . This difficulty might explain why boundary losses have been, to the best of our knowledge, completely avoided in the context of deep segmentation networks.

Our boundary loss is inspired from discrete (graph-based) optimization techniques for computing gradient flows of curve evolution (Boykov et al., 2006). Similarly to our problem, curve evolution methods require a measure for evaluating boundary changes (or variations). Consider the

---

2. We focus on two-region segmentation to simplify the presentation. However, our formulation extends to the multi-region case in a straightforward manner.

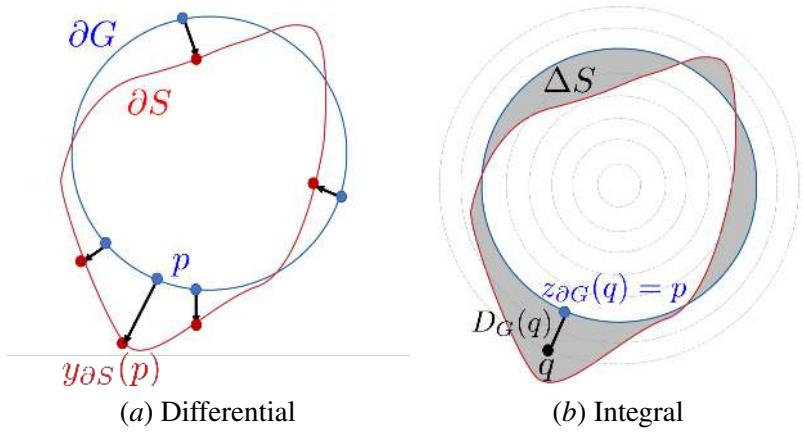


Figure 2: The relationship between *differential* and *integral* approaches for evaluating boundary change (variation).

following non-symmetric  $L_2$  distance on the space of shapes, which evaluates the change between two nearby boundaries  $\partial S$  and  $\partial G$  (Boykov et al., 2006):

$$\text{Dist}(\partial G, \partial S) = \int_{\partial G} \|y_{\partial S}(p) - p\|^2 dp \quad (2)$$

where  $p \in \Omega$  is a point on boundary  $\partial G$  and  $y_{\partial S}(p)$  denotes the corresponding point on boundary  $\partial S$ , along the direction normal to  $\partial G$ , i.e.,  $y_{\partial S}(p)$  is the intersection of  $\partial S$  and the line that is normal to  $\partial G$  at  $p$  (See Fig. 2.a for an illustration).  $\|\cdot\|$  denotes the  $L_2$  norm. In fact, this *differential* framework for evaluating boundary change is in line with standard variational curve evolution methods (Mitiche and Ben Ayed, 2011), which compute the motion of each point  $p$  on the evolving curve as a velocity along the normal to the curve at point  $p$ . Similarly to any contour distance invoking directly points on contour  $\partial S$ , expression (2) cannot be used directly as a loss for  $\partial S = \partial S_\theta$ . However, it is easy to show that the differential boundary variation in (2) can be expressed using an *integral* approach (Boykov et al., 2006), which avoids completely local differential computations involving contour points and represents boundary change as a regional integral:

$$\text{Dist}(\partial G, \partial S) = 2 \int_{\Delta S} D_G(q) dq \quad (3)$$

where  $\Delta S$  denotes the region between the two contours and  $D_G : \Omega \rightarrow \mathbb{R}^+$  is a *distance map* with respect to boundary  $\partial G$ , i.e.,  $D_G(q)$  evaluates the distance between point  $q \in \Omega$  and the nearest point  $z_{\partial G}(q)$  on contour  $\partial G$ :  $D_G(q) = \|q - z_{\partial G}(q)\|$ . Fig. 2.b illustrates this integral framework for evaluating the boundary distance in Eq. (2). To show that Eq. (3) holds, it suffices to notice that integrating the distance map  $2D_G(q)$  over the normal segment connecting a point  $p$  on  $\partial G$  and  $y_{\partial S}(p)$  yields  $\|y_{\partial S}(p) - p\|^2$ . This follows directly from a variable change:

$$\int_p^{y_{\partial S}(p)} 2D_G(q) dq = \int_0^{\|y_{\partial S}(p) - p\|} 2D_G dD_G = \|y_{\partial S}(p) - p\|^2$$

Thus, from Eq. (3), the non-symmetric  $L_2$  distance between contours in Eq. (2) can be expressed as a sum of regional integrals based on a *level set* representation of boundary  $\partial G$ :

$$\frac{1}{2}\text{Dist}(\partial G, \partial S) = \int_S \phi_G(q)dq - \int_G \phi_G(q)dq = \int_{\Omega} \phi_G(q)s(q)dq - \int_{\Omega} \phi_G(q)g(q)dq \quad (4)$$

where  $s : \Omega \rightarrow \{0, 1\}$  is binary indicator function of region  $S$ :  $s(q) = 1$  if  $q \in S$  belongs to the target and 0 otherwise.  $\phi_G : \Omega \rightarrow \mathbb{R}$  denotes the level set representation of boundary  $\partial G$ :  $\phi_G(q) = -D_G(q)$  if  $q \in G$  and  $\phi_G(q) = D_G(q)$  otherwise. Now, for  $S = S_{\theta}$ , i.e., replacing binary variables  $s(q)$  in Eq. (4) by the softmax probability outputs of the network  $s_{\theta}(q)$ , we obtain the following boundary loss which, up to a constant independent of  $\theta$ , approximates boundary distance  $\text{Dist}(\partial G, \partial S_{\theta})$ :

$$\mathcal{L}_B(\theta) = \int_{\Omega} \phi_G(q)s_{\theta}(q)dq \quad (5)$$

Notice that we omitted the last term in Eq. (4) as it is independent of network parameters. The level set function  $\phi_G$  is pre-computed directly from the ground-truth region  $G$ . In practice, our boundary loss in Eq. (5) is the sum of linear functions of the regional softmax probability outputs of the network. Therefore, it can be easily combined with standard regional losses and implemented with any existing deep network architecture for N-D segmentation. In the experiments, we will use our boundary loss in conjunction with the regional generalized Dice loss:

$$\alpha \mathcal{L}_{GD}(\theta) + (1 - \alpha) \mathcal{L}_B(\theta) \quad (6)$$

Finally, it is worth noting that our boundary loss uses ground-truth boundary information via pre-computed level-set function  $\phi_G(q)$ , which encodes the distance between each point  $q$  and  $\partial G$ . In Eq. (5), the softmax for each point  $q$  is weighted by the distance function. Such distance-to-boundary information is omitted in widely used regional losses, where all the points within a given region are treated equally, independently of their distances from the boundary.

### 3. Experiments

#### 3.1. Datasets

To evaluate the proposed boundary loss, we selected two challenging brain lesion segmentation tasks, each corresponding to highly unbalanced classes.

**ISLES:** The training dataset provided by the ISLES organizers is composed of 94 ischemic stroke lesion multi-modal scans. In our experiments, we split this dataset into training and validation sets containing 74 and 20 examples, respectively. Each scan contains Diffusion maps (DWI) and Perfusion maps (CBF, MTT, CBV, Tmax and CTP source data), as well as the manual ground-truth segmentation. More details can be found in the ISLES website<sup>3</sup>.

**WMH:** The public dataset of the White Matter Hyperintensities (WMH)<sup>4</sup> MICCAI 2017 challenge contains 60 3D T1-weighted scans and 2D multi-slice FLAIR acquired from multiple vendors and scanners in three different hospitals. In addition, the ground truth for the 60 scans is provided. From the whole set, 50 scans were used for training, and the remaining 10 for validation.

3. <http://www.isles-challenge.org>

4. <http://wmh.isi.uu.nl>

### 3.2. Implementation details

**Data pre-processing.** While the scans are provided as 3D images, we process them as a stack of independent 2D images, which are fed into the network. In fact, the scans in some datasets, such as ISLES, contain between 2 and 16 slices, making them ill-suited for 3D convolutions in those cases. The scans were normalized between 0 and 1 before being saved as a set of 2D matrices, and re-scaled to  $256 \times 256$  pixels if needed. When several modalities were available, all of them were concatenated before being used as input to the network. We did not use any data augmentation in our experiments.

**Architecture and training.** We employed UNet (Ronneberger et al., 2015) as deep learning architecture in our experiments. To train our model, we employed Adam optimizer, with a learning rate of 0.001 and a batch size equal to 8. The learning rate is halved if the validation performances do not improve during 20 epochs. We did not use early stopping.

To compute the level set function  $\phi_G$  in Eq. (5), we used standard SciPy functions<sup>5</sup>. Note that, for slices containing only the background region, we used a zero-distance map, assuming that the GDL is sufficient in those cases. Furthermore, during training, the value of  $\alpha$  in Eq. (6) was initially set to 1, and decreased by 0.01 after each epoch, following a simple scheduling strategy, until it reached the value of 0.01. In this way, we give more importance to the regional loss term at the beginning while gradually increasing the impact of the boundary loss term. We empirically found that this simple scheduling strategy was less sensitive to the choice of  $\alpha$  while giving consistently similar or better results than a constant value. In addition, we evaluated the performance when the boundary loss is the only objective, i.e.,  $\alpha = 0$ .

For our implementation, we used PyTorch (Paszke et al., 2017), and ran the experiments on a machine equipped with an NVIDIA GTX 1080 Ti GPU with 11GBs of memory. Our code (data pre-processing, training and testing scripts) is publicly available<sup>6</sup>.

**Evaluation.** For evaluation purposes, we employ the common Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) metrics.

### 3.3. Results

**Quantitative evaluation.** Table 1 reports the DSC and HD performance for our experiments using GDL alone and the loss we proposed in Eq. (6) on the ISLES and WMH datasets. Adding our boundary loss term to the GDL consistently improves segmentation performance, which is reflected in significantly higher DSC and HD values. While this boundary loss term brings a DSC improvement of around 2% on the WMH dataset, it achieves 8% better DSC on the ISLES segmentation task. The same trend is observed with the HD metric, where the gain is larger on the ISLES dataset than on WMH.

Using the boundary loss alone does not yield the same competitive results as a joint loss (i.e., boundary and region), making the network collapse quickly into empty foreground regions, i.e., softmax predictions close to zero<sup>7</sup>. We believe that this is due to the following technical facts. In theory, the global optimum of the boundary loss corresponds to a negative value, as a perfect overlap

---

5. [https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.ndimage.morphology.distance\\_transform\\_edt.html](https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.ndimage.morphology.distance_transform_edt.html)

6. <https://github.com/LIVIAETS/surface-loss>

7. For this reason, Hausdorff distances are not reported for this case, as it would be meaningless.

Table 1: DSC and HD values achieved on the validation subset. The values represent the mean performance (and standard deviation) of 2 runs for each setting.

Loss	ISLES		WMH	
	DSC	HD (mm)	DSC	HD (mm)
$\mathcal{L}_B$	0.321 (0.000)	NA	0.569 (0.000)	NA
$\mathcal{L}_{GD}$	0.575 (0.028)	4.009 (0.016)	0.727 (0.006)	1.045 (0.014)
$\mathcal{L}_{GD} + \mathcal{L}_B$	<b>0.656 (0.023)</b>	<b>3.562 (0.009)</b>	<b>0.748 (0.005)</b>	<b>0.987 (0.010)</b>

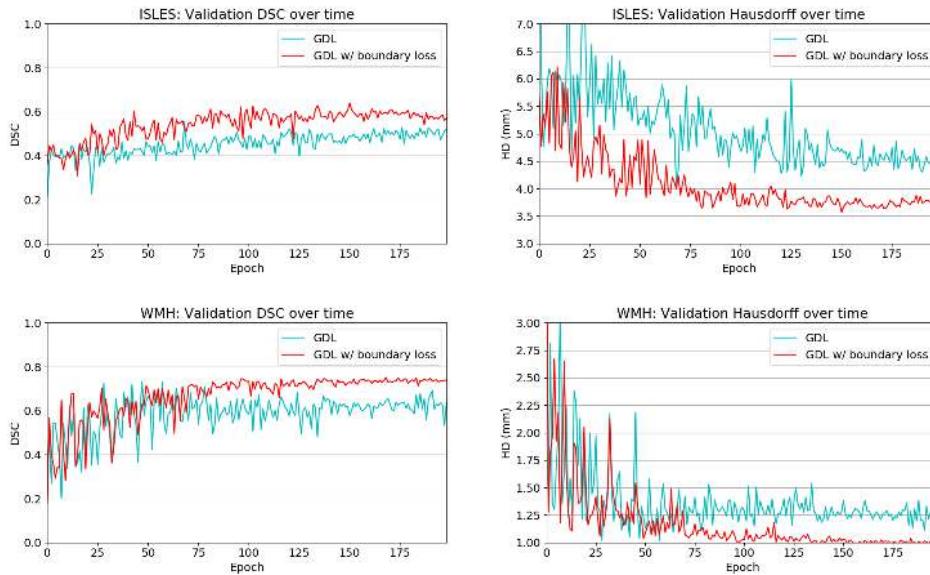


Figure 3: Evolution of DSC and HD values on the validation subset when training on ISLES and WMH dataset. The blue curve shows the performance of the network trained using the GDL loss, while the red curve represents the optimization process with the GDL + our proposed boundary loss term.

sums only over the negative values of the distance map. In this case, the softmax probabilities correspond to a non-empty foreground. However, an empty foreground (null values of the softmax probabilities almost everywhere) corresponds to low gradients. Therefore, this trivial solution is close a local minimum or a saddle point. This is not the case when we use our boundary loss in conjunction with a regional loss, which guides the training during the first epochs and avoids getting stuck in such a trivial solution. The scheduling method then increases the weight of the boundary loss, with the latter becoming very dominant towards the end of the training process. This behaviour of boundary terms is conceptually similar to the behaviour of classical and popular contour-based energies for level set segmentation, e.g., geodesic active contours (Caselles et al., 1997), which also require additional regional terms to avoid trivial solutions (i.e., empty foreground regions).

The learning curves depicted in Figure 3 show the gap in performance between the GDL alone and the GDL augmented with our boundary loss, with the difference becoming significant at convergence. In addition to outperforming GDL, we can also observe that the boundary loss term helps stabilizing the training process, yielding a much smoother curve as the network training converges. This behaviour is consistent for both metrics and both dataset, which clearly shows the benefits of employing the proposed boundary loss term.

**Qualitative evaluation.** Qualitative results are depicted in Fig. 4. Inspecting these results visually, we can observe that there are two major types of improvements when employing the proposed boundary loss. First, as the methods based on DSC losses, such as GDL, do not use spatial information, prediction errors are treated equally. This means that the errors for pixels/voxels in an already detected object have the same importance as the errors produced in completely missed objects. On the contrary, as our boundary loss is based on the distance map from the ground-truth boundary  $\partial G$ , it will penalize much more such cases, helping to recover small and far regions. This effect is best illustrated in Fig. 1 and Fig. 4 (third row). False positives (first row in Fig. 4) will be far away from the closest foreground, getting a much higher penalty than with the GDL alone. This helps in reducing the number of false positives.

**Computational complexity.** It is worth mentioning that, as the proposed boundary loss term involves an element-wise product between two matrices – i.e., the pre-computed level-set function  $\phi_G$  and the softmax output  $s_\theta(p)$  – the complexity that it adds is negligible.

## 4. Conclusion

We proposed a boundary loss term that can be easily combined with standard regional losses to tackle the segmentation task in highly unbalanced scenarios. Furthermore, the proposed term can be implemented in any existing deep network architecture and for any N-D segmentation problem. Our experiments on two challenging and highly unbalanced datasets demonstrated the effectiveness of including the proposed boundary loss term during training. It consistently improved the performance, with a large margin on one data set, and enhanced training stability. Even though we limited the experiments to 2-D segmentation problems, the proposed framework can be trivially extended to 3-D, which could further improve the performance of deep networks, as more context is analyzed.

## Acknowledgments

This work is supported by the National Science and Engineering Research Council of Canada (NSERC), discovery grant program, and by the ETS Research Chair on Artificial Intelligence in Medical Imaging.

## References

- Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention U-Net for lesion segmentation. *arXiv preprint arXiv:1810.07842*, 2018.
- Yuri Boykov, Vladimir Kolmogorov, Daniel Cremers, and Andrew Delong. An integral solution to surface evolution PDEs via geo-cuts. In *European Conference on Computer Vision*, pages 409–422. Springer, 2006.

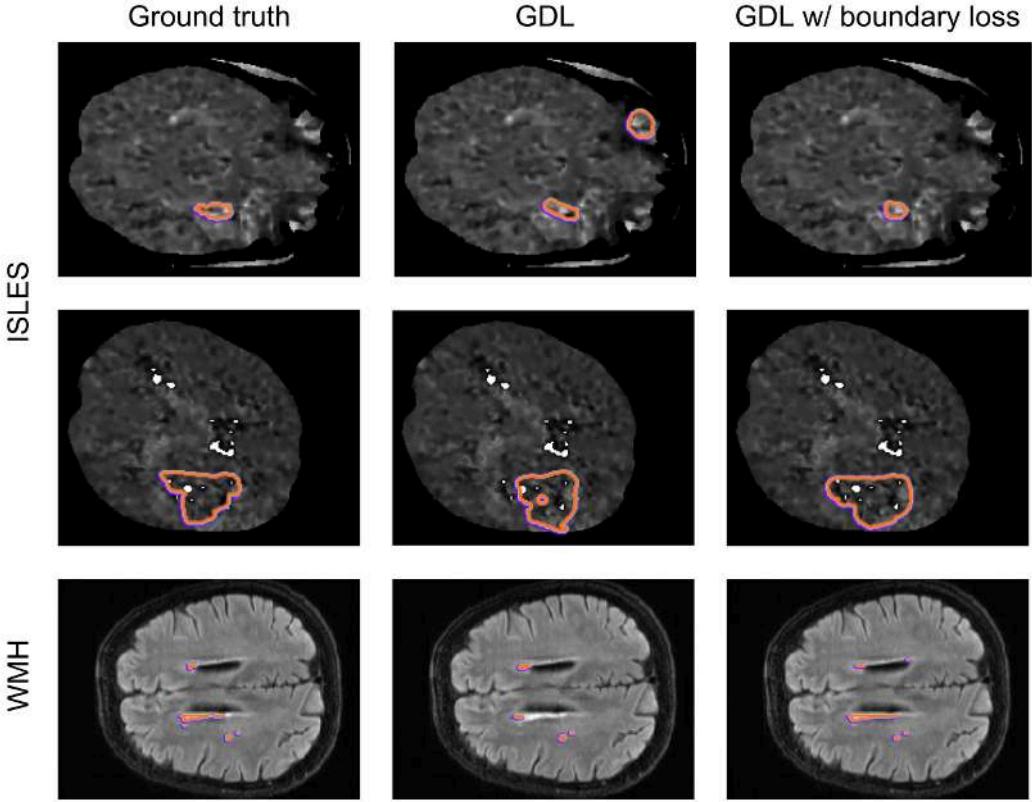


Figure 4: Visual comparison on two different datasets from the validation set.

Tom Brosch, Youngjin Yoo, Lisa YW Tang, David KB Li, Anthony Traboulsee, and Roger Tam. Deep convolutional encoder networks for multiple sclerosis lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–11. Springer, 2015.

V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22:61–79, 1997.

Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, 170:456–470, 2018.

Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

- Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. Deep learning applications in medical image analysis. *IEEE Access*, 6:9375–9389, 2018.
- Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- Geert Litjens, Thiss Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.
- Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.
- Fausto Milletari, Seyed-Ahmad Ahmadi, Christine Kroll, Annika Plate, Verena Rozanski, Julian Maiostre, Johannes Levin, Olaf Dietrich, Birgit Ertl-Wagner, Kai Bötz, et al. Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. *Computer Vision and Image Understanding*, 164:92–102, 2017.
- Amar Mitiche and Ismail Ben Ayed. *Variational and level set methods in image segmentation*. Springer, 2011.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop Autodiff Submission*, 2017.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 379–387. Springer, 2017.
- Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.
- Carole H Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248. Springer, 2017.
- Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramí-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.

Ken CL Wong, Mehdi Moradi, Hui Tang, and Tanveer Syeda-Mahmood. 3D segmentation with exponential logarithmic loss for highly unbalanced object sizes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 612–619. Springer, 2018.

Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng-Ann Heng. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3D MR images. In *AAAI*, pages 66–72, 2017.

# Neural Processes Mixed-Effect Models for Deep Normative Modeling of Clinical Neuroimaging Data

Seyed Mostafa Kia<sup>1,2</sup>

S.KIA@DONDERS.RU.NL

Andre F. Marquand<sup>1,2,3</sup>

A.MARQUAND@DONDERS.RU.NL

<sup>1</sup> Department of Cognitive Neuroscience, Radboud University Medical Center, Nijmegen, Netherlands

<sup>2</sup> Donders Institute for Brain Cognition and Behaviour, Nijmegen, Netherlands

<sup>3</sup> Department of Neuroimaging, Institute of Psychiatry, King's College, London, United Kingdom

## Abstract

Normative modeling has recently been introduced as a promising approach for modeling variation of neuroimaging measures across individuals in order to derive biomarkers of psychiatric disorders. Current implementations rely on Gaussian process regression, which provides coherent estimates of uncertainty needed for the method but also suffers from drawbacks including poor scaling to large datasets and a reliance on fixed parametric kernels. In this paper, we propose a deep normative modeling framework based on neural processes (NPs) to solve these problems. To achieve this, we define a stochastic process formulation for mixed-effect models and show how NPs can be adopted for spatially structured mixed-effect modeling of neuroimaging data. This enables us to learn optimal feature representations and covariance structure for the random-effect and noise via global latent variables. In this scheme, predictive uncertainty can be approximated by sampling from the distribution of these global latent variables. On a publicly available clinical fMRI dataset, we compare the novelty detection performance of multivariate normative models estimated by the proposed NP approach to a baseline multi-task Gaussian process regression approach and show substantial improvements for certain diagnostic problems.

**Keywords:** Neural Processes, Mixed-Effect Modeling, Deep Learning, Neuroimaging.

## 1. Introduction

Recently, there has been great interest in applying machine learning to neuroimaging in order to find structural or functional biomarkers for brain disorders (Bzdok and Meyer-Lindenberg, 2018). Such biomarkers can potentially be used for diagnosis or predicting treatment outcome in the spirit of *precision medicine* (Mirnezami et al., 2012). In psychiatry, this is very challenging because clinical groups are highly heterogeneous in terms of symptoms and underlying biology (Kapur et al., 2012). However, most common analysis approaches ignore such heterogeneity and, in a case-control setting consider groups as distinct entities (Foulkes and Blakemore, 2018), where subjects are simply labeled as ‘patients’ or ‘controls’. Supervised machine learning methods have been widely used in such settings but their accuracy is limited by the heterogeneity within each disorder (Wolfers et al., 2015).

Normative modeling (Marquand et al., 2016) is an emerging approach to address this challenge that has shown significant promise in multiple clinical settings (Wolfers et al., 2018; Zabihi et al., 2018; Wolfers et al., 2019). Normative modeling involves estimating variation across the population in terms of mappings between clinically relevant covariates (e.g., age, cognitive scores) and biology (e.g., neuroimages). This is analogous to the use of ‘growth charts’ in pediatric medicine to map

variation in height or weight as a function of age. Currently, this is implemented using probabilistic regression methods that provide estimates of predictive uncertainty which map variation across the population. Deviations from the resulting *normative* model can then be interpreted as subject-specific biomarkers for brain disorders. For example, these can be used in a novelty detection setting for predicting diagnosis in an *unsupervised* fashion (Kia and Marquand, 2018; Kia et al., 2018).

Accurate quantification of uncertainty is crucial for normative modeling. In the original framework (Marquand et al., 2016), Gaussian process regression (Williams and Rasmussen, 1996) (GPR) was the central tool used to regress neuroimaging measures from clinical covariates. GPR is appealing because it estimates a distribution over functions, providing coherent estimates of uncertainty to map population variation. However GPR also has limitations: it is computationally prohibitive for large datasets and relies on predefined kernels with restricted functional form. Moreover, in the original implementation, brain measures were regressed independently (i.e., in a mass-univariate manner), which does not capitalize on the rich spatial structure of neuroimaging data. This last problem can be addressed by using multi-task GPR (MT-GPR) (Bonilla et al., 2008) to jointly predict multiple brain measurements. However, applying MT-GPR to neuroimaging data is very computationally demanding because of the need to invert large covariance matrices across both space and subjects. Recently, a combination of low-rank approximations and Kronecker algebra was proposed to scale MT-GPR to whole brain neuroimaging data (Kia and Marquand, 2018; Kia et al., 2018), which reduces the computational complexity with respect to the number of tasks by one order of magnitude. However, this comes with restrictive assumptions that the spatial structures of the signal and noise can be expressed by sets of orthogonal basis functions. Furthermore, its times complexity still remains cubic with the number of samples which is not appropriate for applications on large clinical cohorts.

Neural processes (NP) (Garnelo et al., 2018a,b) are latent variable models that bring all the advantages of deep learning (e.g., representation learning and computationally efficient training and prediction) to the stochastic process framework and can address the problems described above. In the NP framework, a distribution over functions is modeled by learning an approximation to a stochastic process. Here, we present an application of NP to multivariate normative modeling of clinical neuroimaging data. This provides three advantages: i) like GPR, NP provides the necessary estimates of predictive uncertainty at test time; ii) similar to MT-GPR, it provides the possibility of learning structured variation; and iii) unlike alternatives, it is computationally scalable without restrictive assumptions on the orthogonality of lower dimensional representations of data. To this end, we make four contributions: i) in a tensor Gaussian predictive process (TGPP) framework (Kia et al., 2018), we formally define mixed-effect models of neuroimaging data (Friston et al., 1999) as stochastic processes; ii) we show how NP can be employed for mixed-effect modeling; iii) we use the resulting NP-based mixed-effect model to estimate a normative model of a clinical functional magnetic resonance imaging (fMRI) dataset; iv) we provide an example application of the proposed *deep* normative modeling for detecting psychiatric disorders in a novelty detection setting. Our experimental results show that the proposed method more accurately identifies ADHD patients from healthy individuals compared to the GP-based normative modeling.

## 2. Methods

In this text, we use respectively calligraphic capital letters,  $\mathcal{A}$ , boldface capital letters,  $\mathbf{A}$ , and capital letters,  $A$ , to denote tensors, matrices, and scalars. We use  $\times_1$  to denote 1st-mode tensor product.

We denote the vertical vector which results from collapsing the entries of a tensor  $\mathcal{A}$  into a vector with  $\text{vec}(\mathcal{A})$ . Notation  $|.|$  is accordingly used to represents the determinant of a matrix or the size of a set.

## 2.1. Mixed-Effect Modeling of MRIs in the TGPP Framework

Consider a neuroimaging study with  $N$  subjects and let  $\mathbf{X} \in \mathbb{R}^{N \times D}$  denote the design matrix of  $D$  covariates of interest for  $N$  subjects. Let  $\mathcal{Y} \in \mathbb{R}^{N \times T_1 \times T_2 \times T_3}$  represent a 4-order tensor of MRI data for corresponding  $N$  subjects with respectively  $T_1$ ,  $T_2$ , and  $T_3$  voxels in  $x$ ,  $y$ , and  $z$  axes. In the normative modeling setting, we are interested in finding the function  $f : \mathbf{X} \rightarrow \mathcal{Y}$ . Adopting the tensor Gaussian predictive process (TGPP) (Kia et al., 2018) for structured multi-way mixed-effect modeling of MRI data, we have:

$$\mathcal{Y} = f(\mathbf{X}) = \mathbf{X} \times_1 \mathcal{A} + \mathcal{Z} + \mathcal{E}, \quad (1)$$

where  $\mathcal{A} \in \mathbb{R}^{D \times T_1 \times T_2 \times T_3}$  represents the *fixed-effect* across subjects that contains regression coefficients estimated by solving the following linear equations:

$$\hat{\mathcal{Y}}[:, i, j, k] = \mathbf{X} \mathcal{A}[:, i, j, k], \quad \text{for } i = 1, \dots, T_1; \quad j = 1, \dots, T_2; \quad k = 1, \dots, T_3. \quad (2)$$

In Equation (1),  $\mathcal{Z} \in \mathbb{R}^{N \times T_1 \times T_2 \times T_3}$  is the *random-effect* that characterizes the spatially structured joint variations from the fixed-effect across individuals in different dimensions of MRIs; and  $\mathcal{E} \in \mathbb{R}^{N \times T_1 \times T_2 \times T_3}$  is heteroscedastic noise. Assuming a tensor-variate normal distribution for  $\mathcal{Y}$  and a zero-mean tensor-variate normal distribution for  $\mathcal{Z} + \mathcal{E}$ , we have:

$$p(\mathbf{X}, \mathcal{Y}) = \mathcal{T}\mathcal{N}(\hat{\mathcal{Y}}, \mathbf{S}) = \frac{\exp(-\frac{1}{2} \text{vec}(\mathcal{Y} - \hat{\mathcal{Y}})^\top \mathbf{S}^{-1} \text{vec}(\mathcal{Y} - \hat{\mathcal{Y}}))}{\sqrt{(2\pi)^{NT} |\mathbf{S}|^{NT}}}, \quad (3)$$

where  $\mathbf{S} \in \mathbb{R}^{NT \times NT}$  ( $T = T_1 \times T_2 \times T_3$ ) is the covariance matrix of  $\mathcal{Z} + \mathcal{E}$ . Intuitively, the distribution of the mixed-effect in the joint hypercubic space of clinical covariates and neuroimaging measures can be described as a multi-dimensional Gaussian distribution with  $\text{vec}(\hat{\mathcal{Y}})$  and  $\mathbf{S}$  respectively serving as its mean and covariance.

## 2.2. Mixed-Effect Models of MRI Data as Stochastic Processes

The primary aim of this section is to formally define the structured mixed-effect model in Equation (1) as stochastic process. This will provide the ingredients to employ NP for learning characteristics of the covariance matrix of the random-effect and noise in Equation (3), i.e.,  $\mathbf{S}$ .

Let  $(\Omega, \Phi, \rho)$  represent a complete probability space (see Oksendal (2003) or Appendix B for definitions) where  $\Omega$  is a set of clinical covariates and their corresponding neuroimaging measures pairs for  $N$  subjects (i.e.,  $|\Omega| = N$ ) and  $\Phi$  is a  $\sigma$ -algebra on  $\Omega$  that contains all possible subsets of  $\Omega$ . Here,  $\rho : \Phi \rightarrow [0, 1]$  represents a probability measure that quantifies the probability of occurrence for any entry in  $\Phi$ . In this setting, each mixed-effect function  $f_i$  estimated on the  $i$ th entry of  $\Phi$  is a random variable, i.e., a  $\Phi$ -measurable function from  $\Omega$  to a Borel set in  $\mathbb{R}^{N \times T_1 \times T_2 \times T_3}$ . Therefore, parametrizing  $f_i$  on different subsets of  $\Omega$ ; and considering the exchangeability and consistency properties of mixed-effect models (McCullagh, 2005; Nie and Yang, 2005),  $\mathcal{Y}_i = f_i(\mathbf{X}_i) |_{i=1}^{|\Phi|}$  can be defined as stochastic processes (Garnelo et al., 2018b). As a corollary, for the  $i$ th entry in  $\Phi$ ,

$\phi_i = (\mathbf{X}_i, \mathcal{Y}_i) \subset \Omega$  with  $|\phi_i| = N_i < N$ , the joint distribution  $p(\mathbf{X}_i, \mathcal{Y}_i)$  can be considered as a marginal for a higher-dimensional joint distribution in Equation (3). We exploit this property to frame the problem of mixed-effect modeling in the neural processes framework (Garnelo et al., 2018b). To this end, given a particular realization of the mixed-effect stochastic process  $f_i$ , the joint distribution in Equation (3) can be rewritten as:

$$p(\mathbf{X}, \mathcal{Y}) = \sum_{i=1}^{|\Phi|} p(f_i) \mathcal{T}\mathcal{N}(\mathcal{Y} | f_i, \mathbf{S}). \quad (4)$$

In an NP paradigm (see Appendix A.1 for background information on NP), we parametrize the integration over all  $f_i(\mathbf{X})$  on a lower dimensional ( $Q \ll T$ ) Gaussian distributed global latent variable  $\mathbf{Z} \in \mathbb{R}^{N \times Q} \sim \mathcal{N}(\mu, \Sigma)$  where  $f(\mathbf{X}) = g(\mathbf{X}, \mathbf{Z})$ , resulting the following generative model:

$$p(\mathbf{Z}, \mathcal{Y} | \mathbf{X}) = p(\mathbf{Z}) \mathcal{T}\mathcal{N}(\mathcal{Y} | g(\mathbf{X}, \mathbf{Z}), \mathbf{S}) , \quad (5)$$

where  $g(\mathbf{X}, \mathbf{Z})$  is a deep neural network that learns the behavior of the mixed-effect model in an amortized variational inference regime (Kingma and Welling, 2013; Gershman and Goodman, 2014). To this end, following the procedure proposed by Garnelo et al. (2018b) the first challenge is to induce stochasticity, for which we need to define ‘context’ and ‘target’ points. While target points refer to full available information (e.g., all pixels in an image), the context points are intended to represent some partial information about the target function (e.g., a subset of pixels in an image). In this work, in order to adapt the NP for the mixed-effect modeling, we advance the concepts of context/target points (Garnelo et al., 2018b) to context/target functions (see Section 5.2 for discussion). The idea is to reduce the difference between the distribution of random context functions from the target function by minimizing their Kullback-Leibler (KL) divergence in the latent space. In our application in order to learn the distribution of the mixed-effect model in Equation (1), i.e., target function, we propose to use the estimated  $\hat{\mathcal{Y}}_C \in \mathbb{R}^{N \times M \times T_1 \times T_2 \times T_3}$  (using Equation (2)) on  $M$  randomly drawn subsets of the training set as context functions. Then, using the actual corresponding neuroimaging training samples as target functions, the following evidence lower-bound should be optimized:

$$\log p(\mathcal{Y} | \mathbf{X}, \hat{\mathcal{Y}}_C) \geq \mathbb{E}_{q(\mathbf{Z} | \mathbf{X}, \mathcal{Y})} \left[ \log p(\mathcal{Y} | \mathbf{Z}, \mathbf{X}) + \log \frac{q(\mathbf{Z} | \mathbf{X}, \hat{\mathcal{Y}}_C)}{q(\mathbf{Z} | \mathbf{X}, \mathcal{Y})} \right] , \quad (6)$$

where  $q(\mathbf{Z} | \mathbf{X}, \mathcal{Y})$  is the variational posterior of the global latent variable that is parametrized on an encoder  $h(\mathbf{X}, \hat{\mathcal{Y}}_C)$ . In fact in this setting, each context function is a linear component of the target function that roughly approximates a stochastic process  $f_i$ . Having enough samples of context functions, large enough  $M$ , we expect the distribution of context functions to get rich enough to explain non-linear characteristics of the target function (i.e., the mixed-effect  $f_i$ ). Figure 1 shows a simplified illustration of this scenario in a 2D space where fitting enough linear models on subsets of noisy observations provides an estimation of the distribution of a non-linear target function. Furthermore, by minimizing the KL term in Equation (6), it is expected that the global latent variable  $\mathbf{Z}$  will learn characteristics of the variance structure of the random-effect and noise terms (the diagonal elements of  $\mathbf{S}$ ) from the difference between the context and target functions (recall that  $\mathcal{Y} - \hat{\mathcal{Y}} = \mathcal{Z} + \mathcal{E}$ ).

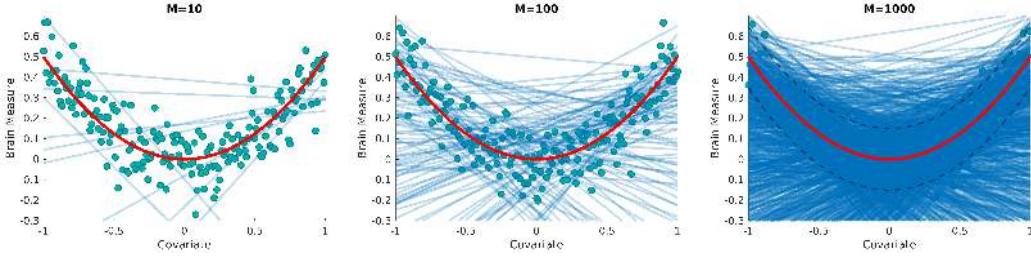


Figure 1: A schematic illustration on approximating the distribution of a non-linear target function (red curve), e.g., a mixed-effect, from the distribution of linear context functions (blue lines), e.g., fixed-effects, which are fitted on  $M$  random subsets of noisy observations (circles).

### 2.3. Deep Normative Modeling using Neural Processes

Using NP in the TGPP framework brings all the advantages of deep learning methods (e.g., representation learning from structured data and computational efficiency) for modeling the multi-way structured variation in neuroimaging data. It has been shown that modeling such structured variation provides the possibility of accurate unsupervised stratification of psychiatric patients in the normative modeling paradigm (Kia and Marquand, 2018; Kia et al., 2018). To this end, here we introduce *deep normative modeling*, which utilizes an NP-based mixed-effect modeling and involves following three steps:

- 1. Encoding phase:** where an encoder  $h(\mathbf{X}, \hat{\mathcal{Y}}_C)$  is learned to transfer the covariates,  $\mathbf{X}$ , and the estimated fixed-effects on  $M$  randomly drawn samples from the training set,  $\hat{\mathcal{Y}}_C$ , to the parameters of the global latent variable  $\mathbf{Z}$ . Here, to preserve the 3D MRIs structure in the TGPP framework, we propose to use 3D-convolutional neural network (3D-CNN) layers to first transfer the  $\hat{\mathcal{Y}}_C$  to a lower dimensional representation of neuroimages  $\mathbf{R}_{\hat{\mathcal{Y}}} \in \mathbb{R}^{N \times T'}$ . Note that using a CNN architecture in NP complicates fusing  $\mathbf{X}$  with  $\hat{\mathcal{Y}}_C$  in the encoder. When using fully-connected layers in the encoder (for example in Garnelo et al. (2018b)), this fusion is simply performed by concatenation. However, considering inherent structural differences between  $\mathbf{X}$  and  $\hat{\mathcal{Y}}_C$  the concatenation is impossible when using a CNN architecture. Therefore, this concatenation is performed in the latent output space  $\mathbf{R}_{\hat{\mathcal{Y}}}$  (see Section 5.3 for discussion on its advantages). Then, fully connected (FC) layers can be used to derive a latent representation in the joint space of clinical covariates ( $\mathbf{X}$ ) and neuroimages,  $\mathbf{R} \in \mathbb{R}^{N \times T''}$ . It is worthwhile to emphasize that in this architecture, the aggregation across  $M$  context functions is implicitly done by the 3D-CNN layers as they are considered as  $M$  input channels to the CNN. Finally, two separate FC layers are used to transfer  $\mathbf{R}$  to the means ( $\mu_{\mathbf{Z}} \in \mathbb{R}^{N \times Q}$ ) and standard deviations ( $\sigma_{\mathbf{Z}} \in \mathbb{R}^{N \times Q}$ ) of  $\mathbf{Z}$ .
- 2. Decoding phase:** where a decoder  $g(\mathbf{X}, \mathbf{Z})$  is learned to transfer back the joint covariates-latent space to the neuroimaging data  $\mathcal{Y}$ . Fully connected and 3D inverse CNN (3D-ICNN) layers can be accordingly used to reconstruct MRIs in the original space.

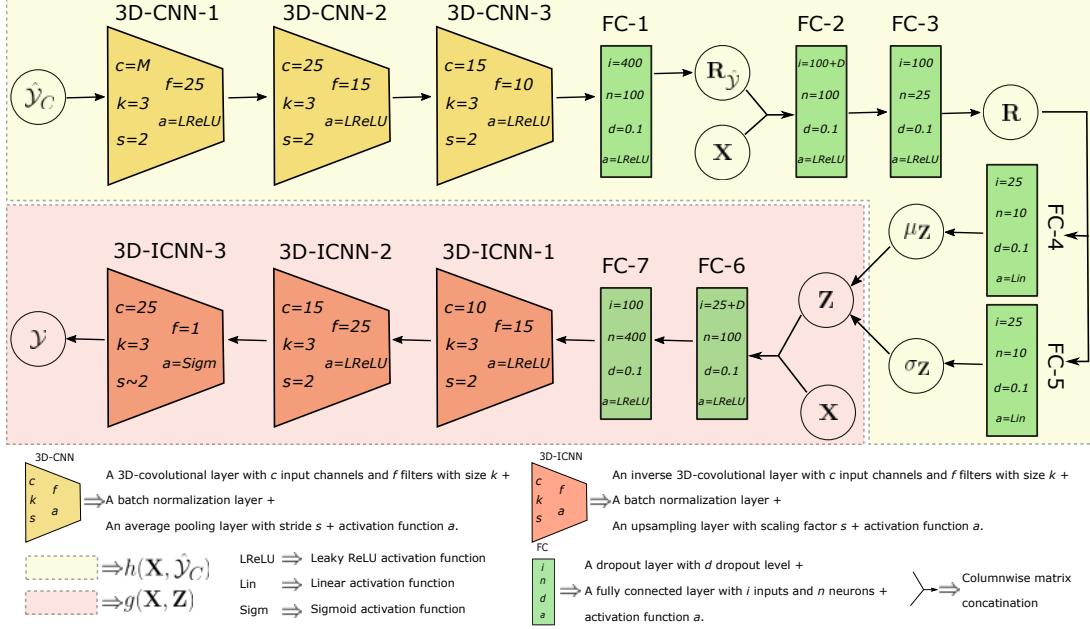


Figure 2: An example NP architecture for mixed-effect modeling of MRIs.

3. **Normative modeling:** let  $\mathcal{Y}^* \in \mathbb{R}^{N^* \times T_1 \times T_2 \times T_3}$  to represent the reconstructed neuroimaging data by the decoder  $g(\mathbf{X}^*, \mathbf{Z})$  on  $N^*$  test samples. Following Marquand et al. (2016) (see Appendix A.2 for details), we then compute statistical maps describing the deviation for each individual subject from the normative model, referred to as normative probability maps (NPMs), denoted by  $\mathcal{N} \in \mathbb{R}^{N^* \times T_1 \times T_2 \times T_3}$  where  $\mathcal{N} = (\mathcal{Y} - \mathcal{Y}^*) / \sqrt{\mathcal{S}}$ . Here,  $\mathcal{S}$  represents the sum of epistemic and aleatoric uncertainties, which respectively describe uncertainty about the true model parameters and inherent variation in the data (Kendall and Gal, 2017). To be able to calculate the epistemic uncertainty in our NP model, we keep the dropout layers active at test time (Gal and Ghahramani, 2016). In the context of mixed-effect modeling of neuroimaging data (in Equation (1)), the aleatoric uncertainty is the byproduct of two factors: i) the across-subject variability which is captured via the covariance of the random-effect  $\mathbf{Z}$ ; and ii) noise in the data which is captured via covariance of  $\mathcal{E}$ . In the proposed NP framework, these two sources of uncertainties are learned from data and are summarized in the distribution of the global latent variable  $\mathbf{Z}$ . Therefore, given a test example of clinical covariates  $\mathbf{x}^* \in \mathbf{X}^*$ , we calculate the associated aleatoric uncertainty by sampling from the distribution of  $\mathbf{Z}$ .

### 3. Experimental Materials and Setup

In our experiments, we use the response inhibition (i.e., ‘stop signal’) task from the UCLA Consortium for Neuropsychiatric Phenomics dataset (Poldrack et al., 2016). Specifically, we use the ‘Go’ contrast volumes derived from the pipeline in Gorgolewski et al. (2017).<sup>1</sup> The data consist of

1. Available at <https://openfmri.org/dataset/ds000030/>.

119 healthy subjects; and 49, 39, and 48 individuals with schizophrenia (SCHZ), attention deficit hyperactivity disorder (ADHD), and bipolar disorder (BIPL), respectively. We cropped the volumes to the minimal bounding-box of  $49 \times 61 \times 40$  voxels ( $T_1 = 49, T_2 = 61, T_3 = 40, T = 119560$ ). In order to accommodate the optimization scheme in Equation (6) for fMRI data, the values of voxels are independently projected to the uniform  $[0, 1]$  interval using a robust quantile transformation. For clinical covariates, we use 11 factors of Barratt impulsiveness scores (Patton et al., 1995) ( $D = 11$ ) as impulsivity is a well-known feature for multiple psychiatric disorders and is implicated in response inhibition (Moeller et al., 2001).

We use three layers of 3D-CNNs followed by an FC layer to project  $\hat{\mathcal{Y}}_C$  to  $\mathbf{R}_{\hat{\mathcal{Y}}}$ . In each CNN layer, we alternate a 3D-convolutional layer, a batch normalization layer (Ioffe and Szegedy, 2015), an average pooling layer, and a leaky ReLU activation function (Xu et al., 2015) (with negative slope of 0.01). Then, two FC layers are used to transfer the merged  $\mathbf{R}_{\hat{\mathcal{Y}}}$  and  $\mathbf{X}$  to the middle joint representation  $\mathbf{R}$ . A similar reverse architecture is used for the decoder  $g(\mathbf{X}, \mathbf{Z})$  to transfer back the  $\mathbf{Z}$  to  $\mathcal{Y}$  space. Figure 2 depicts a schematic of the employed NP architecture with detailed hyperparameter descriptions. Due to the small sample size and illustrative purpose of our experiments, we did not optimize the architecture and its hyperparameters (e.g., number of layers, number and the size of filters, number of neurons, etc.). The ADAM optimizer (Kingma and Ba, 2014) with decreasing learning rate (from  $10^{-2}$  to  $10^{-5}$ ) is used for optimization in 100 epochs.

We compare the normative models derived by NP and scalable multi-task Gaussian process tensor regression (sMT-GPTR) (Kia et al., 2018), in terms of their accuracy in detecting healthy subjects from patients.<sup>2</sup> In the sMT-GPTR case, we set the number of basis functions across xyz dimensions of data 5, 10 and 3, 5 for the signal and noise, respectively (as they produced the best results in the original study). We evaluate normative modeling accuracy in a novelty detection scenario where we first train a model on a random subset of majority healthy subjects (75 healthy, 5 SCHZ, 5 ADHD, and 5 BIPL) and then calculate NPMs on a test set of remaining healthy subjects and patients.  $\sim 16\%$  of cases are included in the training set in order to seemingly simulate the average prevalence of general mental disorders in a cohort (Consortium, 2004). We emphasize that the model has no access to the diagnostic labels during the training phase and thus our novelty detection approach is completely unsupervised. As in Marquand et al. (2016), we use extreme value statistics to provide a statistical model for the deviations (see Appendix A.3 for more details). Specifically, we use a block-maximum approach on the top 1% values in NPMs and fit these to a generalized extreme value distribution (GEVD) (Davison and Huser, 2015). Then for a given test sample and given the shape parameter of GEVD, we compute the value of the cumulative distribution function of GEVD as the probability of that sample being an abnormal sample (Roberts, 2000). Given these probabilities and actual labels, we evaluate the area under the receiver operating characteristic curve (AUC) to measure the performance of the model. All steps (random sampling, modeling, and evaluation) are repeated 10 times in order to estimate the fluctuations of models trained on different training sets. In all these experiments, ordinary least squares are used to estimate the fixed-effect (Equation (2)) on bootstrapped subsets of the training set.<sup>3</sup>

---

2. The implementation for sMT-GPTR is available at <https://github.com/smchia/MTNorm>.

3. The scripts for experiments are available at <https://github.com/smchia/DNM>.

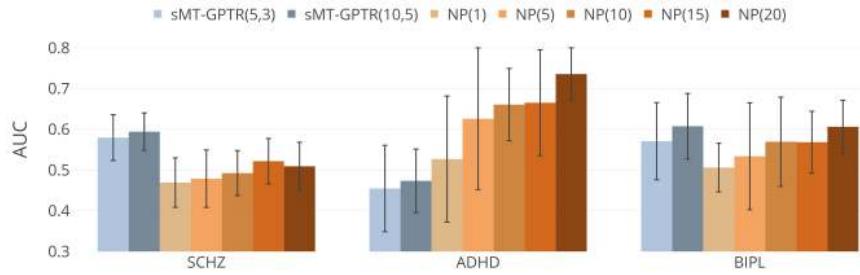


Figure 3: Comparison between novelty detection performances of normative models derived by sMT-GPTR (with different number of bases for signal and noise) and NP (with different  $M$ ).

#### 4. Results

Figure 3 compares the AUC of normative models derived by sMT-GPTR and NP. While sMT-GPTR shows slightly better performance in detecting SCHZ patients, NP provides substantially higher accuracy for ADHD cases. The methods perform similarly for BIPL. Considering the fact that these differences in performance are consistent across different model parameters and repetitions, it can be concluded that sMT-GPTR and NP are capturing different characteristics of the underlying biology of impulsivity. Furthermore, the above chance-level detection rates of NP models in ADHD and BIPL confirm a successful application of the proposed NP-based mixed-effect modeling in unsupervised diagnostic prediction. The significance of these results are even more pronounced considering the difficulty of the problem where a supervised support vector machine classifier provides only a chance-level performance in ADHD and BIPL cases ( $\text{SCHZ} = 0.67 \pm 0.07$ ,  $\text{ADHD} = 0.46 \pm 0.03$ ,  $\text{BIPL} = 0.47 \pm 0.06$ ).<sup>4</sup> Another important observation in NP models is the ascending trend of the detection performance as the number of samples from the fixed-effect ( $M$ ) increases. This is compatible with the consistency property of mixed-effects as stochastic processes.

Figure 4(a) depicts the average difference in NPMs of patient groups from the healthy population for NP(20) model (see Appendix C for supplementary results). Different patterns of deviations from one diagnosis to another shed light on their different underlying biological causes. For example, the sign of deviations changes from SCHZ to ADHD patients in many regions. To further explore the link between these deviations and the level of impulsivity, we computed the coefficient of determination ( $R^2$ ) between the average NPMs in 9 anatomical brain areas and the first principal component of covariates across different diagnostic groups (see Figure 4(b)). The results show significantly (Bonferroni corrected F-test p-values) greater association between impulsivity and deviations in temporal lobes in ADHD and SCHZ patients compared to healthy individuals. This observation is compatible with previous research on the structural and functional engagement of temporal lobes in SCHZ and ADHD (Suddath et al., 1989; Kobel et al., 2010).

4. See Kia et al. (2018) for training and evaluation configurations in the supervised setting.

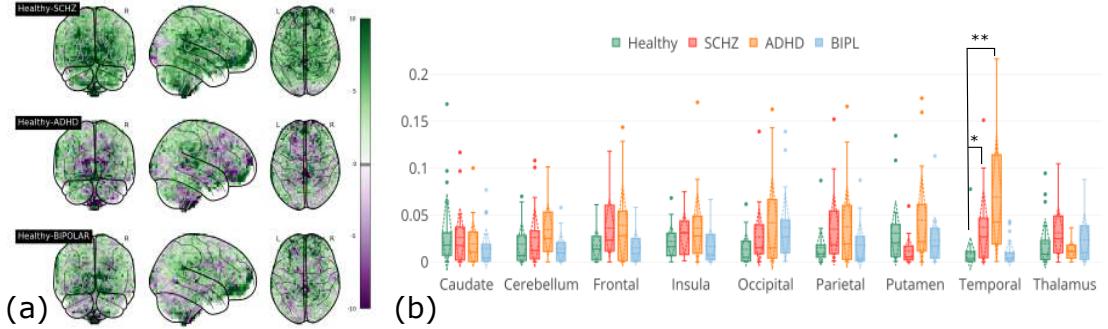


Figure 4: (a) The average difference between NPMs of healthy subjects and patients for NP(20). (b)  $R^2$  between the impulsivity and deviation from the normative model across different anatomical brain areas (\*\*  $p < 0.01$  and \*  $p < 0.1$ ).

## 5. Discussion

### 5.1. Toward Multivariate Normative Modeling on Large Clinical Cohorts

Including spatial information in probabilistic modeling and extending the mass-univariate normative modeling to its multivariate alternative is computationally very expensive. For  $N$  samples and  $T$  tasks, the time complexity of MT-GPR is cubic with respect to the number of samples and tasks,  $\mathcal{O}(N^3 T^3)$ . Many efforts have been devoted in order to reduce the time complexity of MT-GPR for large output spaces (i.e., large  $T$ ) low-rank approximation and properties of Kronecker product (Alvarez and Lawrence, 2009; Stegle et al., 2011; Rakitsch et al., 2013; Kia and Marquand, 2018; Kia et al., 2018). However, for very large sample-size datasets (i.e., for large  $N$  and especially when  $N \gg T$ ), their time complexity still remains cubic with respect to  $N$  that limits their applications in normative modeling on recently available large clinical cohorts (Sudlow et al., 2015) (with  $N \approx 10^4 - 10^5$ ). One possible remedy for this problem is to approximate the posterior distribution of a probabilistic model with hidden variables in the stochastic variational inference framework (Hoffmann et al., 2013). Alvarez et al. (2010) made the first effort in employing variational inference in MT-GPR by introducing the variational inducing kernels that achieves a linear time complexity with respect to  $N$ . Our proposed NP-based normative modeling also employs the variational inference scheme, therefore, its computational complexity remains linear with respect to the number of samples in both training and inference phases (Garnelo et al., 2018b). Furthermore, since the spatial information is incorporated using a CNN architecture, there is no need to compute the inverse covariance matrix for the output space. These two properties make this method very suitable for multivariate normative modeling on large clinical cohorts of high-dimensional neuroimaging data.

### 5.2. From Context/Target Points to Context/Target Functions

In order to learn the joint distribution in Equation (5) over random functions rather than a single function, the evidence lower-bound in Equation (6) is optimized by minimizing the KL divergence between variational posteriors over context and target points. In the original NP framework (Garnelo et al., 2018b), the target points are defined as whole points in the full dataset, while the context

points are defined as a subset of target points that represent a partial knowledge about the full dataset. For example in the case of MNIST dataset, a random subset of pixels in an image can be used for context points. The random selection of pixels in the context points provides the desired stochasticity behavior in the NP framework. In this study, we advance the concepts of context/target points to target/context functions where the idea is to learn the distribution of a non-linear mixed-effect function, i.e., the target function, from a set of linear fixed-effect functions, i.e., context functions, estimated on a random subset of subjects. This alternation is key to learning the characteristics of the variance structure of random-effect and noise via the global latent variable and consequently it is crucial for normative modeling.

### 5.3. Preserving Spatial Structures via Convolutional Neural Processes

In this study, CNN-based architectures are proposed for encoding and decoding operations in NP. This results in two main advantages especially for the applications on neuroimaging data. First, it provides the possibility of preserving spatially-structured information in MRI data. Second, the parameter sharing gifted by CNN substantially reduces the computational costs in training and inference when dealing with very high-dimensional neuroimaging data.

### 5.4. Related Work

[Rad et al. \(2018\)](#) used a convolutional autoencoder for unimodal deep normative modeling of human movements recorded by wearable sensors. They used dropout technique in order to evaluate the parameter uncertainty of the model. Our proposed NP-based approach extends their effort in applying deep architectures to normative modeling from two perspectives: i) it provides the possibility of bimodal normative modeling. This is more appropriate for clinical usages where we are generally interested in interpreting the association between clinical covariates and biological measures ([Marquand et al., 2016](#)); ii) using the fully probabilistic NP regime, we are also capable to evaluate aleatoric uncertainties resulting from individual differences and noise in addition to the epistemic parameter uncertainty.

## 6. Conclusions

In this paper, we proposed a principled approach for estimating spatially structured mixed-effects in neuroimaging data using neural processes. We demonstrated normative modeling as a possible target application for NP-based mixed-effect modeling. Even though the main focus in this study was on neuroimaging data, our contribution in framing the popular mixed-effect modeling as stochastic processes is quite general and opens the door for a wide range of NP applications in different research areas. Moreover, the presented application of NP for deep normative modeling of clinical neuroimaging data brings the advantages of deep neural networks in representation learning to the applications in precision psychiatry. Finally, the computational efficiency of NP in the training and evaluation phases (provided by its reliance on the variational inference) overcomes the lack of computational tractability of the GP-based normative modeling approaches especially when applied to large cohorts of high-dimensional neuroimaging data. For a possible future direction, we consider applying the proposed deep normative modeling approach to a large clinical neuroimaging cohort.

## References

- Mauricio Alvarez and Neil D Lawrence. Sparse convolved Gaussian processes for multi-output regression. In *Advances in neural information processing systems*, pages 57–64, 2009.
- Mauricio A Alvarez, David Luengo, Michalis K Titsias, and Neil D Lawrence. Efficient multi-output Gaussian processes through variational inducing kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 25–32, 2010.
- Edwin V Bonilla, Kian M Chai, and Christopher Williams. Multi-task Gaussian process prediction. In *Advances in neural information processing systems*, pages 153–160, 2008.
- Danilo Bzdok and Andreas Meyer-Lindenberg. Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(3): 223 – 230, 2018. ISSN 2451-9022. doi: <https://doi.org/10.1016/j.bpsc.2017.11.007>. URL <http://www.sciencedirect.com/science/article/pii/S2451902217302069>.
- The WHO World Mental Health Survey Consortium. Prevalence, Severity, and Unmet Need for Treatment of Mental Disorders in the World Health Organization World Mental Health Surveys. *JAMA*, 291(21):2581–2590, 06 2004. ISSN 0098-7484. doi: 10.1001/jama.291.21.2581. URL <https://doi.org/10.1001/jama.291.21.2581>.
- Anthony C Davison and Raphaël Huser. Statistics of extremes. *Annual Review of Statistics and Its Application*, 2(1):203–235, 2015. doi: 10.1146/annurev-statistics-010814-020133. URL <https://doi.org/10.1146/annurev-statistics-010814-020133>.
- Lucy Foulkes and Sarah-Jayne Blakemore. Studying individual differences in human adolescent brain development. *Nature neuroscience*, page 1, 2018.
- Karl J Friston, Andrew P Holmes, CJ Price, C Büchel, and KJ Worsley. Multisubject fMRI Studies and Conjunction Analyses. *NeuroImage*, 10(4):385 – 396, 1999. ISSN 1053-8119. doi: <https://doi.org/10.1006/nimg.1999.0484>. URL <http://www.sciencedirect.com/science/article/pii/S105381199904846>.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL <http://proceedings.mlr.press/v48/gal16.html>.
- Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1704–1713, Stockholm, Sweden, 10–15 Jul 2018a. PMLR. URL <http://proceedings.mlr.press/v80/garnelo18a.html>.
- Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

- Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.
- Krzysztof J Gorgolewski, Joke Durnez, and Russell A Poldrack. Preprocessed consortium for neuropsychiatric phenomics dataset [version 2; referees: 2 approved]. *F1000Research*, 6(1262), 2017. doi: 10.12688/f1000research.11964.2.
- Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347, May 2013. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=2502581.2502622>.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL <http://proceedings.mlr.press/v37/ioffe15.html>.
- Shitij Kapur, Anthony G Phillips, and Thomas R Insel. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular psychiatry*, 17(12):1174, 2012.
- Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, pages 5580–5590, 2017.
- Seyed Mostafa Kia and Andre Marquand. Normative modeling of neuroimaging data using scalable multi-task gaussian processes. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 127–135, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00931-1.
- Seyed Mostafa Kia, Christian F. Backmann, and Andre F. Marquand. Scalable multi-task gaussian process tensor regression for normative modeling of structured variation in neuroimaging data. *arXiv preprint arXiv:1808.00036*, 2018.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Maja Kobel, Nina Bechtel, Karsten Specht, Markus Klarhöfer, Peter Weber, Klaus Scheffler, Klaus Opwis, and Iris-Katharina Penner. Structural and functional imaging approaches in attention deficit/hyperactivity disorder: Does the temporal lobe play a key role? *Psychiatry Research: Neuroimaging*, 183(3):230 – 236, 2010. ISSN 0925-4927. doi: <https://doi.org/10.1016/j.pscychresns.2010.03.010>. URL <http://www.sciencedirect.com/science/article/pii/S0925492710001137>.
- Andre F Marquand, Iead Rezek, Jan Buitelaar, and Christian F Beckmann. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biological psychiatry*, 80(7):552–561, 2016.

- Peter McCullagh. Exchangeability and regression models. *Oxford Statistical Science Series*, 33:89, 2005.
- Reza Mirnezami, Jeremy Nicholson, and Ara Darzi. Preparing for Precision Medicine. *New England Journal of Medicine*, 366(6):489–491, 2012. doi: 10.1056/NEJMp1114866. PMID: 22256780.
- F. Gerard Moeller, Ernest S. Barratt, Donald M. Dougherty, Joy M. Schmitz, and Alan C. Swann. Psychiatric aspects of impulsivity. *American Journal of Psychiatry*, 158(11):1783–1793, 2001. doi: 10.1176/appi.ajp.158.11.1783. URL <https://doi.org/10.1176/appi.ajp.158.11.1783>. PMID: 11691682.
- Lei Nie and Min Yang. Strong consistency of mle in nonlinear mixed-effects models with large cluster size. *Sankhyā: The Indian Journal of Statistics*, pages 736–763, 2005.
- Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Berlin, 2003.
- Jim H. Patton, Matthew S. Stanford, and Ernest S. Barratt. Factor structure of the barratt impulsiveness scale. *Journal of Clinical Psychology*, 51(6):768–774, 1995. doi: 10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-4679%28199511%2951%3A6%3C768%3A%3AAID-JCLP2270510607%3E3.0.CO%3B2-1>.
- Russell A Poldrack, Eliza Congdon, William Triplett, KJ Gorgolewski, KH Karlsgodt, JA Mumford, FW Sabb, NB Freimer, ED London, TD Cannon, et al. A genome-wide examination of neural and cognitive function. *Scientific data*, 3:160110, 2016.
- Nastaran Mohammadian Rad, Twan van Laarhoven, Cesare Furlanello, and Elena Marchiori. Novelty detection using deep normative modeling for imu-based abnormal movement monitoring in parkinson’s disease and autism spectrum disorders. *Sensors*, 18(10), 2018. ISSN 1424-8220. doi: 10.3390/s18103533. URL <http://www.mdpi.com/1424-8220/18/10/3533>.
- Barbara Rakitsch, Christoph Lippert, Karsten Borgwardt, and Oliver Stegle. It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals. In *Advances in neural information processing systems*, pages 1466–1474, 2013.
- S.J. Roberts. Extreme value statistics for novelty detection in biomedical data processing. *IEE Proceedings - Science, Measurement and Technology*, 147:363–367(4), November 2000. ISSN 1350-2344. URL [http://digital-library.theiet.org/content/journals/10.1049/ip-smt\\_20000841](http://digital-library.theiet.org/content/journals/10.1049/ip-smt_20000841).
- Oliver Stegle, Christoph Lippert, Joris M Mooij, Neil D Lawrence, and Karsten M Borgwardt. Efficient inference in matrix-variate gaussian models with iid observation noise. In *Advances in neural information processing systems*, pages 630–638, 2011.
- Richard L Suddath, Manuel F Casanova, Terry E Goldberg, David G Daniel, John R Kelsoe Jr, and Daniel R Weinberger. Temporal lobe pathology in schizophrenia: a quantitative magnetic resonance imaging study. *American Journal of Psychiatry*, 146(4):464–472, 1989. doi: 10.1176/ajp.146.4.464. URL <https://doi.org/10.1176/ajp.146.4.464>. PMID: 2929746.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):1–10, 03 2015. doi: 10.1371/journal.pmed.1001779. URL <https://doi.org/10.1371/journal.pmed.1001779>.

Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for regression. In *Advances in neural information processing systems*, pages 514–520, 1996.

Thomas Wolfers, Jan K. Buitelaar, Christian F. Beckmann, Barbara Franke, and Andre F. Marquand. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience and Biobehavioral Reviews*, 57:328 – 349, 2015. ISSN 0149-7634. doi: <https://doi.org/10.1016/j.neubiorev.2015.08.001>. URL <http://www.sciencedirect.com/science/article/pii/S0149763415002018>.

Thomas Wolfers, Nhat Trung Doan, Tobias Kaufmann, and et al. Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. *JAMA Psychiatry*, 75(11):1146–1155, 2018. doi: 10.1001/jamapsychiatry.2018.2467. URL [+http://dx.doi.org/10.1001/jamapsychiatry.2018.2467](http://dx.doi.org/10.1001/jamapsychiatry.2018.2467).

Thomas Wolfers, Christian F. Beckmann, Martine Hoogman, Jan K. Buitelaar, Barbara Franke, and Andre F. Marquand. Individual differences v. the average patient: mapping the heterogeneity in adhd using normative models. *Psychological Medicine*, page 1–10, 2019. doi: 10.1017/S0033291719000084.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

Mariam Zabihi, Marianne Oldehinkel, Thomas Wolfers, Vincent Frouin, David Goyard, Eva Loth, Tony Charman, Julian Tillmann, Tobias Banaschewski, Guillaume Dumas, Rosemary Holt, Simon Baron-Cohen, Sarah Durston, Sven Bölte, Declan Murphy, Christine Ecker, Jan K. Buitelaar, Christian F. Beckmann, and Andre F. Marquand. Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2018. ISSN 2451-9022. doi: <https://doi.org/10.1016/j.bpsc.2018.11.013>. URL <http://www.sciencedirect.com/science/article/pii/S245190221830329X>.

## Appendix A. Backgrounds

### A.1. Neural Processes

A neural process (NP) (Garnelo et al., 2018b) provides a computational tool to learn the distribution over a set of functions from distributions over a set of datasets  $\Phi$ . Assuming the  $i$ th dataset in  $\Phi$  to contain a set of  $N_i$  input-output pairs  $(\mathbf{X}_i, \mathbf{Y}_i)$  where  $\mathbf{X}_i \in \mathbb{R}^{N_i \times D}$  and  $\mathbf{Y}_i \in \mathbb{R}^{N_i \times T}$  and we have  $f_i : \mathbf{X}_i \rightarrow \mathbf{Y}_i$ . For sake of simplicity, we refer to all  $\mathbf{X}_i$  and  $\mathbf{Y}_i$  in  $\Phi$  as  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively. The goal of NP is to learn the distribution of  $f_i$ s from  $(\mathbf{X}, \mathbf{Y})$  pairs in  $\Phi$  via learning the distribution of a global latent variable  $\mathbf{Z}$  in the variational inference framework. For the generative model of an NP we have:

$$p(\mathbf{Z}, \mathbf{Y} | \mathbf{X}) = p(\mathbf{Z})p(\mathbf{Y} | g(\mathbf{X}, \mathbf{Z})) = \mathcal{N}(\mathbf{Y} | g(\mathbf{X}, \mathbf{Z}), \mathbf{S}) , \quad (7)$$

where  $g(\mathbf{X}, \mathbf{Z})$  is the decoder function and parametrized by a neural network and  $\mathbf{S}$  is the covariance matrix in the output space. Intuitively, the latent variable  $\mathbf{Z}$  is intended to learn the statistical characteristics of the distribution of  $f : \mathbf{X} \rightarrow \mathbf{Y}$ . Then, the following approximation of variational posterior distribution is used in order to perform the approximate inference in NP:

$$q(\mathbf{Z} | \mathbf{X}, \mathbf{Y}) = \mathcal{N}(m(\mathbb{H}(\mathbf{X}, \mathbf{Y})), s(\mathbb{H}(\mathbf{X}, \mathbf{Y}))) , \quad (8)$$

where,  $h(\mathbf{X}, \mathbf{Y})$  is the encoder function that is parametrized on neural network,  $\mathbb{H}$  is the aggregator operator (for example, mean), and  $m(\cdot)$  and  $s(\cdot)$  are neural networks that map the aggregated values to the mean and standard deviation of  $\mathbf{Z}$ . Using the approximate variational posterior distribution in Equation (8), the evidence lower bound (ELBO) on the log marginal likelihood is derived as follows:

$$\log p(\mathbf{Y} | \mathbf{X}) \geq \mathbb{E}_{q(\mathbf{Z} | \mathbf{X}, \mathbf{Y})} \left[ \log p(\mathbf{Y} | \mathbf{Z}, \mathbf{X}) + \log \frac{p(\mathbf{Z})}{q(\mathbf{Z} | \mathbf{X}, \mathbf{Y})} \right] . \quad (9)$$

In the NP framework, in order to learn such a distribution over random functions rather than a single function we need to create a *context* set of  $M$  datasets  $\Lambda \subset \Phi$  each of which containing input-output context pairs  $(\mathbf{X}_\Lambda, \mathbf{Y}_\Lambda)$ . These datasets are intended to represent some partial information about the target function  $f : (\mathbf{X}, \mathbf{Y})$ . Thus, Equation (9) can be rewritten as:

$$\log p(\mathbf{Y} | \mathbf{X}, \mathbf{X}_\Lambda, \mathbf{Y}_\Lambda) \geq \mathbb{E}_{q(\mathbf{Z} | \mathbf{X}, \mathbf{Y})} \left[ \log p(\mathbf{Y} | \mathbf{Z}, \mathbf{X}) + \log \frac{p(\mathbf{Z} | \mathbf{X}_\Lambda, \mathbf{Y}_\Lambda)}{q(\mathbf{Z} | \mathbf{X}, \mathbf{Y})} \right] , \quad (10)$$

where the prior  $p(Z)$  is replaced by the conditional prior  $p(\mathbf{Z} | \mathbf{X}_\Lambda, \mathbf{Y}_\Lambda)$ . Considering the intractability of this conditional prior we can approximate it by  $q(\mathbf{Z} | \mathbf{X}_\Lambda, \mathbf{Y}_\Lambda)$ . Therefore we optimize the following lower-bound in order to learn the distribution of  $\mathbf{Z}$ :

$$\log p(\mathbf{Y} | \mathbf{X}, \mathbf{X}_\Lambda, \mathbf{Y}_\Lambda) \geq \mathbb{E}_{q(\mathbf{Z} | \mathbf{X}, \mathbf{Y})} \left[ \log p(\mathbf{Y} | \mathbf{Z}, \mathbf{X}) + \log \frac{q(\mathbf{Z} | \mathbf{X}_\Lambda, \mathbf{Y}_\Lambda)}{q(\mathbf{Z} | \mathbf{X}, \mathbf{Y})} \right] . \quad (11)$$

### A.2. Normative Modeling

Normative modeling provides a framework for statistical inference on how the biological brain readouts of each individual subject deviate from the norm of a large population (Marquand et al., 2016). Given  $\mathbf{X} \in \mathbb{R}^{N \times D}$  and  $\mathbf{Y} \in \mathbb{R}^{N \times T}$  respectively as matrix of  $D$  clinical covariates and  $T$  biological brain measures for  $N$  subjects, normative modeling is performed in three steps:

1. finding a mapping function  $f : \mathbf{X} \rightarrow \mathbf{Y}$  from clinical covariates to brain readouts. While a wide range of linear and non-linear models can be used for this mapping. However, since computing the normative probability maps (see the next step) is strongly depends on estimating the prediction uncertainties, Bayesian regression approaches are the best candidates for normative modeling.
2. calculating ‘normative probability maps’ (NPMs),  $\mathbf{Z} \in \mathbb{R}^{N \times T}$ , as follows:

$$\mathbf{Z} = \frac{\mathbf{Y} - \hat{\mathbf{Y}}}{\sqrt{\mathbf{S}}} \quad , \quad (12)$$

where  $\hat{\mathbf{Y}}$  and  $\mathbf{S}$  are prediction mean and uncertainty, respectively. NPMs can be used to localize brain-related abnormalities at the single subject level ([Wolfers et al., 2018](#); [Zabihi et al., 2018](#); [Wolfers et al., 2019](#)). To ensure accurate estimation of the NPMs it is important to model different sources of variation in data and model.

3. computing subject-level summary statistics using a block-maximum approach by averaging top 1% values in NPM of each subject. These summary statistics across subjects can be used as inputs to a novelty detection algorithm for diagnosis purposes ([Kia and Marquand, 2018](#); [Kia et al., 2018](#); [Rad et al., 2018](#)).

### A.3. Novelty Detection using Generalized Extreme Value Distribution

According to [Marquand et al. \(2016\)](#), we can fit a generalized extreme value distribution (GEVD) on normative summary statistics across subjects in order to compute the abnormality index for each subject. This abnormality index can be defined as the probability of each sample being an abnormal sample by computing the cumulative distribution function of the fitted GEVD ([Roberts, 2000](#)). For a random variable  $a \in \mathbb{R}$ , the cumulative distribution function of the GEVD is defined as below ([Davison and Huser, 2015](#)):

$$F(a) = \begin{cases} \exp(-[1 + \xi(a - \mu)/\sigma]^{-1/\xi}), & \xi \neq 0 \\ \exp(-\exp(-(a - \mu)/\sigma)), & \xi = 0 \end{cases} \quad (13)$$

$\mu \in \mathbb{R}$  and  $\sigma > 0$  are respectively the location and scale parameters and  $\xi \in \mathbb{R}$  is the shape parameter. Depending on whether  $\xi < 0$ ,  $\xi = 0$ , or  $\xi > 0$  the GEVD follows the special cases of the Weibull, Gumbel, Fréchet distributions, respectively.

## Appendix B. Supplementary Definitions

Here are some complementary definitions from general probability theory to understand better the concepts in Section 2.2. The definitions are restated from [Oksendal \(2003\)](#).

**Definition 1** *If  $\Omega$ - is a given set, then a  $\sigma$ -algebra  $\Phi$  on  $\Omega$  is a family  $\Phi$  of subsets of  $\Omega$ - with the following properties:*

1.  $\emptyset \in \Phi$ ,
2.  $\forall \phi \in \Phi \Rightarrow \phi^C \in \Phi$ , where  $\phi^C$  is the complement set of  $\phi$  in  $\Omega$ ,

$$3. \phi_1, \phi_2, \dots \in \Phi \Rightarrow \bigcup_{i=1}^{\infty} \phi_i \in \Phi.$$

Then, the pair  $(\Omega, \Phi)$  is called a measurable space and the subsets of  $\Omega$  that belong to  $\Phi$  are called  $\Phi$ -measurable sets.

**Definition 2** A probability measure  $\rho$  on a measurable space  $(\Omega, \Phi)$  is defined as a function  $\rho : \Phi \rightarrow [0, 1]$  such that:

1.  $\rho(\emptyset) = 0, \rho(\Omega) = 1,$
2. if  $\phi_1, \phi_2, \dots \in \Phi$  and  $\forall i, \forall j, i \neq j \Rightarrow \phi_i \cap \phi_j = \emptyset$  then  $\rho\left(\bigcup_{i=1}^{\infty} \phi_i\right) = \sum_{i=1}^{\infty} \rho(\phi_i).$

The triple  $(\Omega, \Phi, \rho)$  is called a probability space.

**Definition 3** A probability space  $(\Omega, \Phi, \rho)$  is called a complete probability space if  $\Phi$  contains all subsets  $\Lambda$  of  $\Omega$ - with  $P$ -outer measure zero, i.e.,  $\forall \phi \in \Phi$  with  $\rho(\phi) = 0$  we have  $\forall \lambda \subset \phi \Rightarrow \lambda \in \Phi$ . Any probability space can be made complete simply by adding to  $\Phi$  all sets of outer measure 0 and by extending  $\rho$  accordingly.

**Definition 4** A stochastic process is a parametrized collection of random variables defined on a probability space  $(\Omega, \Phi, \rho)$  and assuming values in  $\mathbb{R}^n$ .

### Appendix C. Supplementary Results

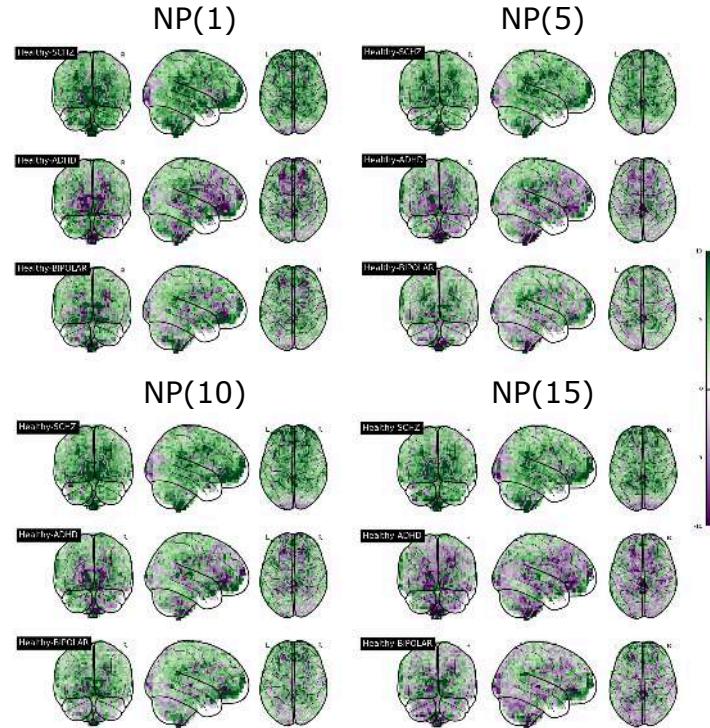


Figure 5: The average difference between NPMs of healthy subjects and patients for NP models with different  $M$ .

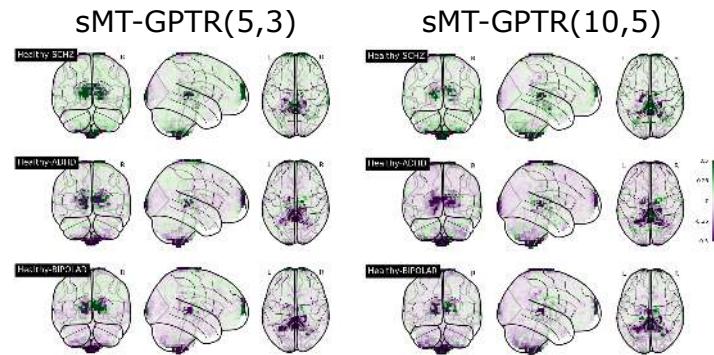


Figure 6: The average difference between NPMs of healthy subjects and patients for sMT-GPTR models with different number of basis functions for the signal and noise.

# Capturing Single-Cell Phenotypic Variation via Unsupervised Representation Learning

**Maxime W. Lafarge<sup>1</sup>**

M.W.LAFARGE@TUE.NL

**Juan C. Caicedo<sup>2</sup>**

JCAICEDO@BROADINSTITUTE.ORG

**Anne E. Carpenter<sup>2</sup>**

ANNE@BROADINSTITUTE.ORG

**Josien P.W. Pluim<sup>1</sup>**

J.PLUIM@TUE.NL

**Shantanu Singh<sup>\*2</sup>**

SHSINGH@BROADINSTITUTE.ORG

**Mitko Veta<sup>\*1</sup>**

M.VETA@TUE.NL

<sup>1</sup> *Medical Image Analysis Group, Department of Biomedical Engineering,*

*Eindhoven University of Technology, Eindhoven, The Netherlands*

<sup>2</sup> *Imaging Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA*

## Abstract

We propose a novel variational autoencoder (VAE) framework for learning representations of cell images for the domain of image-based profiling, important for new therapeutic discovery. Previously, generative adversarial network-based (GAN) approaches were proposed to enable biologists to visualize structural variations in cells that drive differences in populations. However, while the images were realistic, they did not provide direct reconstructions from representations, and their performance in downstream analysis was poor.

We address these limitations in our approach by adding an adversarial-driven similarity constraint applied to the standard VAE framework, and a progressive training procedure that allows higher quality reconstructions than standard VAE’s. The proposed models improve classification accuracy by 22% (to 90%) compared to the best reported GAN model, making it competitive with other models that have higher quality representations, but lack the ability to synthesize images. This provides researchers a new tool to match cellular phenotypes effectively, and also to gain better insight into cellular structure variations that are driving differences between populations of cells.

**Keywords:** Image-based Profiling, Variational Auto-Encoder, Adversarial Training, Biological Interpretability, Fluorescence Microscopy

## 1. Introduction

Microscopy images provide rich information about cell state. Image-based profiling—an approach where images of cells are used as a data source—is a powerful tool with several applications in drug discovery and biomedicine (Caicedo et al., 2016).

Cell samples are treated using chemical or genetic perturbations, then stained using fluorescent markers, and imaged under a microscope. Image-based profiles of these genes or compounds are created by summarizing the single-cell level information extracted from these images. When executed using high-throughput technologies, this framework can be used to generate profiles of tens to hundreds of thousands of perturbations.

---

\* Joint Last Authors

Creating profiles that accurately capture variations in cellular structure is an open problem (Caicedo et al., 2016). Central to this problem is the task of generating representations of single cells, which can then be appropriately summarized into a profile representing the population (e.g. as the mean of the individual cell representations). In recent years, several methods have been proposed for generating single cell representations, spanning both feature engineering approaches (Ljosa et al., 2013), as well as feature learning using deep neural networks (Ando et al., 2017; Caicedo et al., 2018; Pawlowski et al., 2016; Godinez et al., 2017, 2018). While the resulting profiles perform well in downstream analysis, none are able to provide much biological insight into what cellular structure variations are important for discerning phenotypes (i.e. visible appearance). This lack of insight hinders a better understanding of what drives similarities or differences between perturbations.

Recently, generative adversarial networks (GANs) were shown to learn feature representations (Goldsborough et al., 2017), while also to synthesize cell images to help biologists visualize salient phenotypic variations. However, while the images generated were highly realistic, the accuracy of resulting profiles was relatively poor, and a direct reconstruction from the learned representations was not possible.

Here, we propose using an adversarial-driven similarity constraint applied to the standard variational autoencoder (VAE) framework (Kingma and Welling, 2014) that addresses these limitations: (1) VAEs enable direct reconstruction given a feature representation, (2) our proposed model is demonstrably better in learning representations for profiling applications, and (3) our proposed training procedure allows higher quality reconstructions than standard VAEs, making the visualizations comparable with previous GAN models.

By proposing a novel training procedure for learning representations of single cells, we provide researchers a new tool to match cellular phenotypes effectively, and also to gain greater insight into cellular structure variations that are driving differences between populations, offering insights into gene and drug function.

## 2. Related Work

Image-based profiling measures multiple phenotypic features of single cells to characterize the effect of drugs or the function of genes. The phenotypic features can be obtained by engineering representations that capture known relevant properties of cell state, such as cell size. Previous studies using feature engineering approaches demonstrate that profiles generated using standard feature sets in bioimaging software (e.g. CellProfiler (McQuin et al., 2018)) are successful in grouping compounds based on mechanism-of-action (Ljosa et al., 2013; Perlman et al., 2004; Young et al., 2008; Reisen et al., 2015; Ochoa et al., 2015), grouping genes into pathways (Ohya et al., 2005; Fuchs et al., 2010; Rohban et al., 2017), predicting genetic interactions (Horn et al., 2011; Laufer et al., 2013, 2014; Rauscher et al., 2018), and several other applications (Caicedo et al., 2016).

Deep convolutional neural networks (CNN) have been evaluated for computing cellular features using models pretrained on natural images. A deep metric network trained on a large collection of consumer images was evaluated (Ando et al., 2017) for predicting mechanism-of-action in the *BBBC021* benchmark dataset (used in this paper), as were CNNs trained on the *ImageNet* dataset (Pawlowski et al., 2016). Both gave competitive results without requiring cell segmentation or image preprocessing.

Representations can be learned directly from biological images. Multiple instance learning and supervised learning using mechanism-of-action labels directly have been used to train neural networks that process full images without segmentation (Godinez et al., 2017; Kraus et al., 2016). Given that ground truth labels are rarely available for training in high-throughput projects, other strategies that require less supervision have also been explored. Weakly supervised learning using treatment replicates has been proposed to learn single-cell feature embeddings for profiling (Caicedo et al., 2018), and a similar technique has been developed for full fields of view (Godinez et al., 2018).

However, these approaches encode cellular features without an explicit mechanism for interpreting phenotypic variations, a major limitation for many applications in biology. Goldsborough et al. (2017) proposed to tackle this problem using the *CytoGAN* model to generate explanatory visualizations of cell variations between two treatments, but the models do not allow direct reconstructions, and have relatively poor classification accuracy on at least one benchmark dataset (*BBBC021*).

Unlike GAN models, autoencoder (AE) models are optimized to learn embeddings that can directly produce good reconstructions, and were successfully applied on cell images (Ruan and Murphy, 2018; Johnson et al., 2017). In particular, the variational autoencoder (VAE) framework (Kingma and Welling, 2014) implies a constraint on the embedding that produces some desired properties: smooth embedding interpolation (Johnson et al., 2017) and disentanglement of generative factors (Higgins et al., 2016). To improve the limited reconstruction quality of standard VAE models, some methods involving adversarial training were proposed (Larsen et al., 2016; Donahue et al., 2017; Dumoulin et al., 2016; Rosca et al., 2017). Here we propose to follow the concept proposed by Larsen et al. (2016) to address the requirements of the cell profiling pipeline (Caicedo et al., 2017), while allowing high-quality reconstructions from the embeddings.

### 3. Material and Methods

#### 3.1. Datasets

We use the *BBBC021* dataset, a popular benchmark for image-based profiling that has been adopted in several studies, mostly for evaluating assignment of chemicals to mechanisms-of-action using the leave-one-compound-out evaluation protocol (Ljosa et al., 2013). The dataset is from a high-throughput experiment performed in multi-well plates; each plate has 96 wells, and in each well, a sample of cells has been treated with a compound at a specific concentration.

The subset used in all profiling experiments, including ours, has 103 unique treatment conditions (i.e. compounds at a specific concentration) representing 12 mechanisms-of-action (Ljosa et al., 2012). After treatment with a given compound, the cells were stained using fluorescent markers for DNA, F-Actin and  $\beta$ -Tubulin and imaged under a microscope, capturing four 3-channel images for each well, and approximately one million cells across the entire dataset. These channels are stacked and treated as RGB images by mapping DNA  $\mapsto$  R,  $\beta$ -Tubulin  $\mapsto$  G, F-Actin  $\mapsto$  B.

#### 3.2. The VAE framework

In this study, we are interested in methods that can directly generate low-dimensional embeddings  $\mathbf{z}$  and reconstructions  $\tilde{\mathbf{x}}$  of given input images  $\mathbf{x}$ . Therefore we chose the VAE framework as a baseline (Kingma and Welling, 2014). VAE models consist of an encoder convolutional neural network (CNN) that models an approximation of the posterior  $q_\phi(\mathbf{z}|\mathbf{x})$  on the latent  $\mathbf{z}$ , parameterized by  $\phi$ , and a decoder CNN that models the likelihood of the data  $p_\theta(\mathbf{x}|\mathbf{z})$ , parameterized by  $\theta$ .

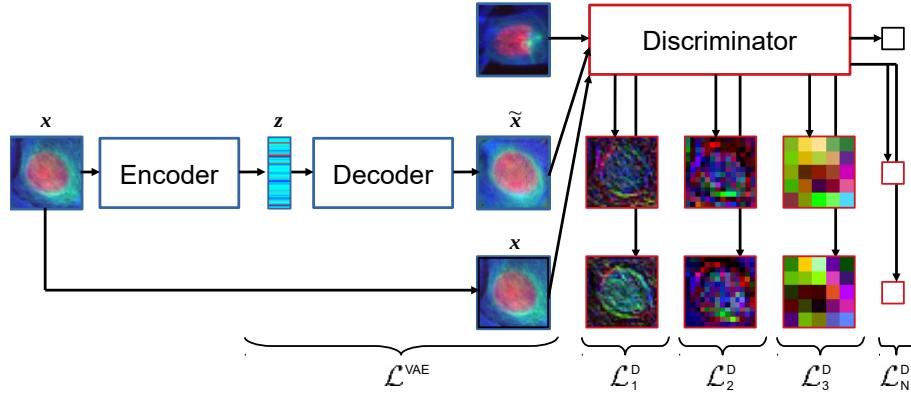


Figure 1: Flowchart of CNN models. The auto-encoder (blue-framed components) describes the original VAE formulation. The adversarial-driven reconstruction losses are illustrated by the representation learned by the discriminator (red-framed images).

The model is then optimized by maximizing a lower bound on the marginal log likelihood of the data  $\mathcal{L}^{\text{VAE}}(\mathbf{x}, \mathbf{z}; \theta, \phi) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \cdot D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}))$ , with  $p(\mathbf{z})$  a defined prior distribution to constrain the embeddings,  $D_{KL}$  the Kullback-Leibler divergence, and  $\beta$  an hyper-parameter controlling the strength of this constraint (Kingma and Welling, 2014; Higgins et al., 2016).

### 3.3. Transition from Pixel-Wise to Adversarial-Driven Reconstructions

The limited reconstruction quality of standard VAE models can be explained by the pixel-wise reconstruction objective related to the Gaussian observation process modeled by  $p_\theta(\mathbf{x}|\mathbf{z})$  (Larsen et al., 2016; Mathieu et al., 2016).

**Learned similarity** As proposed by Larsen et al. (2016), we define a discriminator CNN  $\mathcal{D}$  with parameters  $\chi$  that is trained to classify real images  $\mathbf{x}$  from independent reconstructions  $\tilde{\mathbf{x}}$ . The discriminator outputs the probability for the input to originate from the distribution of real images and is optimized via minimization of the binary cross-entropy.

The activations resulting from the hidden layers of the discriminator  $D_i(\mathbf{x})$  and  $D_i(\tilde{\mathbf{x}})$  are used as additional, synthetic Gaussian observations, with  $i$  the layer indices. These observations are drawn from  $p_\chi(D_i(\mathbf{x})|\mathbf{z})$ , modeled as normal distributions with means  $D_i(\tilde{\mathbf{x}})$  and identity covariances. We thus define additional reconstruction losses  $\mathcal{L}_i^D(\mathbf{x}, \mathbf{z}; \theta, \phi, \chi) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\chi(D_i(\mathbf{x})|\mathbf{z})]$  for every hidden layer  $i$  of the discriminator. Figure 1 illustrates how the different losses of the framework arise in the full model pipeline.

**Progressive Training** We conjecture that the reconstruction term in  $\mathcal{L}^{\text{VAE}}$  should not be discarded and that the additional losses  $\mathcal{L}_i^D$  can be all used to compensate the limited reconstruction ability induced by  $\mathcal{L}^{\text{VAE}}$ , as opposed to the formulation of Larsen et al. (2016). Therefore, we propose to use  $\mathcal{L}^{\text{VAE}} + \mathcal{L}^D$  as the full objective for the encoder and decoder with  $\mathcal{L}^D = \sum_i \gamma_i \cdot \mathcal{L}_i^D$ , and  $(\gamma_i)$  a set of parameters to control the contribution of each reconstruction loss.

For stability purposes when dealing with adversarial training [Karras et al. \(2018\)](#), we chose to define the  $\gamma_i(t)$  as a function of the iteration step  $t$ . By defining  $\gamma_i(t) = \min(1, \max(0, t/T - i))$ , we induce a progressive training procedure, such that the abstraction levels of the discriminator contribute sequentially to  $\mathcal{L}^D$ .  $T$  is thus the hyper-parameter defining the period between two losses  $\mathcal{L}_i^D$  and  $\mathcal{L}_{i+1}^D$  to contribute to the final objective.

### 3.4. Model Architectures

The encoder takes image patches of size  $68 \times 68$  as input, and estimates the mean and standard deviation of the Gaussian posterior, that allow sampling an embedding of size 256 using the reparameterization trick [Kingma and Welling \(2014\)](#).

The encoder, decoder and discriminator have four convolution layers with filters of size  $5 \times 5$ , with an additional  $1 \times 1$  layer for the last layer of the decoder and an additional fully connected layer for the discriminator. Leaky Rectified Linear Units (coefficient 0.01) and max-pooling/up-sampling layers were used throughout the CNNs, except for the last layer of the discriminator, which is activated by a sigmoid.

The decoder is a mirrored version of the encoder, by using transposed convolutions followed by  $2 \times 2$  up-sampling layers. Batch normalization (BN) layers were used throughout the CNNs, and the BN moments for the discriminator were computed only using batches balanced with input reconstructions and independent real images. The implementation of the model is available at <https://github.com/tueimage/cytoVAE>.

## 4. Experiments and Results

### 4.1. Experiments

We investigated three variations of the proposed model for comparison purposes. We trained standard AE and VAE models by setting the parameter  $\beta$  to 0.0 and 1.0 respectively while excluding  $\mathcal{L}^D$  from the full objective. The proposed model (VAE+) was trained using the full objective (see Sect. 3.3),  $\beta$  was set to 2.0 to compensate the additional reconstruction losses, and  $T$  was set to 2500 iterations.

Mini-batches were built by sampling a random image patch from each treatment of the dataset. Every channel of the image patches was normalized by its maximum intensity. Two independent mini-batches  $B_1$  and  $B_2$  were used at every iteration:  $B_1$  was used to compute  $\mathcal{L}^{\text{VAE}}$  through the encoder-decoder,  $B_2$  was paired with the reconstructions of  $B_1$  to train the discriminator. Finally,  $B_1$  and its reconstructions were used to compute  $\mathcal{L}^D$ .

We used the *Adam* optimizer to train the encoder and decoder (learning rate 0.001; momentum 0.9), and Stochastic Gradient Descent with momentum (learning rate 0.01; momentum 0.9) to train the discriminator. All the convolutional weights were regularized with weight decay (coefficient 0.0001). Training was stopped after 40,000 iterations.

### 4.2. Creating Profiles and Classifying Compounds

Given images of cells treated with a compound (the input), the challenge in the *BBBC021* dataset is to predict the mechanism-of-action (the label) of the compound. Centers of each cell were precomputed using CellProfiler ([McQuin et al., 2018](#)) and were used to extract patches. Representations of these patches were generated using the trained models. Given a representation per cell, a profile

Table 1: Classification Accuracy of the compared models. Mean result  $\pm$  standard deviation across 3 repeated experiments with random initialization and random input sampling. The numbers in bold indicate the method-summarization combination that was best performing for each hold-out procedure (NSC and NSCB).

	Method	Mean	Mean +Whitened	Mean+S.D.	Mean+S.D. +Whitened
NSC	VAE+	<b>90.6</b> $\pm$ 1.5	90.3 $\pm$ 1.0	<b>92.2</b> $\pm$ 1.7	<b>92.9</b> $\pm$ 2.4
	VAE	83.5 $\pm$ 1.0	80.6 $\pm$ 4.4	90.9 $\pm$ 1.1	87.1 $\pm$ 0.6
	AE	87.6 $\pm$ 2.0	<b>92.2</b> $\pm$ 1.0	90.3 $\pm$ 0.0	92.5 $\pm$ 0.6
	<i>Ando et al.</i>	N.A	96.0	N.A	N.A
	<i>Singh et al.</i>	90.0	N.A	N.A	N.A
NSCB	VAE+	71.0 $\pm$ 1.2	76.1 $\pm$ 1.1	72.5 $\pm$ 2.3	<b>82.2</b> $\pm$ 2.6
	VAE	68.8 $\pm$ 0.6	69.9 $\pm$ 5.1	74.6 $\pm$ 0.6	71.0 $\pm$ 0.6
	AE	<b>75.0</b> $\pm$ 2.0	<b>79.0</b> $\pm$ 0.6	<b>76.8</b> $\pm$ 0.7	80.8 $\pm$ 1.7
	<i>Ando et al.</i>	N.A	95.0	N.A	N.A
	<i>Singh et al.</i>	85.0	N.A	N.A	N.A

for each well was computed as the average of all the cells in that well. Next, the profile for each unique treatment (a compound at a specific concentration) was computed by computing the median of all wells with that treatment. Treatments were classified using 1-nearest-neighbors, using one of two hold-out procedures as proposed by [Ando et al. \(2017\)](#): (1) Not-Same-Compound (NSC), where all profiles of the same compound (regardless of concentration) were held out, and (2) Not-Same-Compound-and-Batch (NSCB), where in addition to NSC constraints, profiles from the same experimental batch were held out.

NSCB indicates how sensitive the profiling method is to variations across experimental batches; better NSCB performance indicates better resilience to batch variations. [Ando et al. \(2017\)](#) transformed the profiles on a given plate using a whitening transform learned from the control wells on that plate, which improved NSCB performance; we tested this procedure (indicated by “Whitened” in Table 1). Further, [Rohban et al. \(2017\)](#) created profiles by summarizing using standard deviations as well as means; we tested this approach (indicated by “Mean+S.D.” in Table 1).

### 4.3. Results

#### 4.3.1. CLASSIFICATION PERFORMANCES

The proposed VAE model (VAE+ in Table 1) significantly outperforms the best GAN-based models (68% NSC; NSCB unavailable), which is the only model to our knowledge that can provide reconstructions. Further, whitening consistently improves accuracy across all configurations where mean is the summary statistic, and for some where both mean and S.D. are used as summary statistics. The VAE+ model, with mean + S.D. summaries followed by whitening (last column) performs similarly to the best performing classical approach (90% NSC; 85% NSCB [Singh et al. \(2014\)](#)). While none of these models, including VAE+, achieve classification performance as high as the best per-

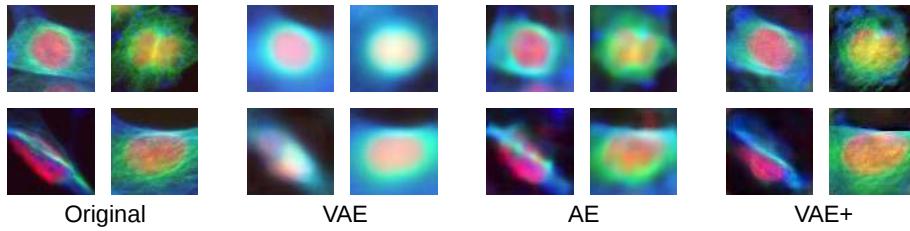


Figure 2: Comparison between original images of four randomly sampled single cells, and their reconstructions produced via different auto-encoders.

forming deep-learning-based model (96% NSC; 95% NSCB [Ando et al. \(2017\)](#)), they nonetheless provide valuable insight (discussed below) into the variations in cellular morphologies that underlie the similarities and differences between the treatment conditions. Finally, we observe that the AE model implemented performed very similarly to VAE+. The VAE+ reconstructions are however superior to AE, making the former overall better suited for profiling applications.

#### 4.3.2. VISUALIZING STRUCTURAL VARIATIONS IN CELL PHENOTYPES

The proposed VAE+ model produces the most realistic images (Figure 2); both AE and VAE images are consistently blurrier than VAE+ images. Similar to [Goldsborough et al. \(2017\)](#), we assessed the quality of reconstructed images by presenting three expert biologists with 50 real cell images and 50 cells reconstructed using VAE+. The cells were balanced across the available treatments, including controls and the biologists were blinded with respect to this treatment information. Images were randomly shuffled and presented to experts to assess whether each cell was real or synthetic. On average, 40.7% of the time the synthetic cells were realistic enough to deceive the experts into labeling them as real, compared to 30% previously reported with GANs ([Goldsborough et al., 2017](#)).

The ability to interpolate between real cells from different treatment conditions and produce realistic images is powerful tool to visualize how a compound affects cellular structure (Figure 3). Compounds from different mechanisms induce visually distinct phenotypes. Interpolating between a control cell and a treated cell presents a hypothetical path in phenotypic space that the cell may have taken to arrive at the observed (target) state. Verifying these hypotheses would require further followup experiments. Regardless, these visualizations give valuable insight into how each compound is affecting cellular structure. For instance, an actin disrupting chemical (*cytochalasin D*) appears to make the cells smaller, with both actin and tubulin condensing more tightly and symmetrically around the nucleus. A cholesterol lowering chemical (*simvastatin*) has a similar effect but makes the tubulin more asymmetric. Both results match expectations and inspection of real images.

However, we noticed one interesting anomaly when exploring a case where VAE+ correctly classified a drug and AE did not (recall their overall classification accuracies across all classes are similar (Table 1)). For the drug *Nocodazole*, a known microtubule destabilizer, AE yields a blurry reconstruction of tubulin while VAE+ yields a more accurate texture (Figure 4). Upon inspection of randomly sampled cell images, however, it becomes clear that neither representation is able to capture the distinctive fragmented nucleus phenotype caused in some cells by *Nocodazole*. We suspect that the selection of the target cell is thus a crucial choice in the proposed strategy, particularly when a population of cells shows two very distinct types of appearances.

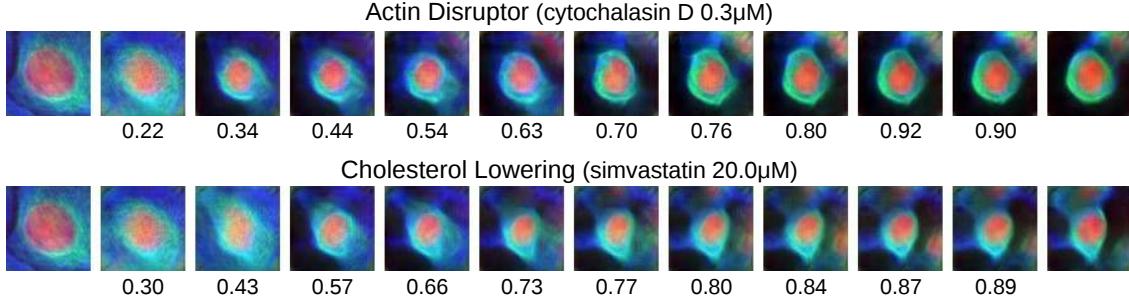


Figure 3: Translation in VAE+ latent space of a control cell (left) to target cells (right) corresponding to compounds with different mechanisms-of-action. The target cell is the one closest to the mean of the compound. Each interpolation step is a shift of features with highest absolute difference w.r.t. the target features. Cosine similarity between the embedding of an image and its target is shown below each.

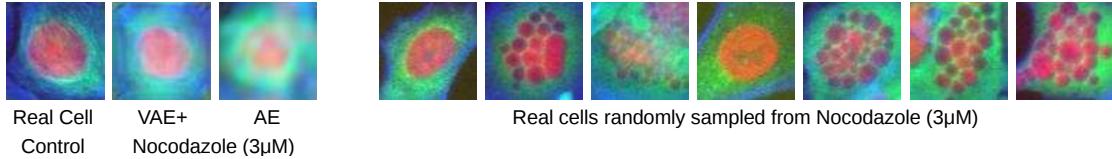


Figure 4: VAE+ captures  $\beta$ -tubulin structure better (less blurry) and correctly identifies *Nocodazole* as a microtubule destabilizer. AE incorrectly classifies it as an actin disruptor. However, neither captures the fragmented nucleus phenotype seen in a fraction of cells’ real images (right).

## 5. Conclusions and Discussion

We proposed an auto-encoding approach competitive with other unsupervised learning approaches while overcoming the challenge of high quality reconstructions.

We introduced adversarial-driven representation learning for the problem of image-based profiling using a straightforward extension of the VAE framework, by proposing a generic method inline with the work of [Larsen et al. \(2016\)](#). Some methods are other plausible solutions for this task, such as Adversarially Learned Inference [Dumoulin et al. \(2016\)](#) and are worth investigating for future work.

The unsupervised training context explains the limited classification performances reported here, but could be improved when combined with more effective approaches (weakly/fully supervised training).

This model offers researchers a powerful tool to probe the structural changes in a cell induced by genetic and chemical perturbations, or even disease states. This is a step towards filling the gap of interpretability in image-based profiling approaches: to reveal not just which perturbations are similar or different, but also to provide clues about the underlying biology that makes them so. We identified room for improvement in capturing phenotypes for very heterogeneous cell populations.

The proposed strategy may be applied to other domains in biomedical imaging that require capturing phenotypic variations, particularly detecting, understanding, and reversing disease.

## Acknowledgments

This work was supported in part by a grant from the COST Action CA15124 (NEUBIAS) and from the National Institutes of Health (MIRA R35 GM122547 to AEC). We thank Peter Goldsborough and Nick Pawlowski for the app to assess real vs. synthetic images, and Drs. Beth Cimini, Minh Doan, and Hamdah Abbasi for doing the assessment. We are grateful to everyone at the Imaging Platform at the Broad for their help throughout this project.

## References

- D Michael Ando, Cory McLean, and Marc Berndl. Improving phenotypic measurements in high-content imaging screens. *bioRxiv*, page 161422, 2017.
- Juan C Caicedo, Shantanu Singh, and Anne E Carpenter. Applications in image-based profiling of perturbations. *Curr. Opin. Biotechnol.*, 39:134–142, June 2016.
- Juan C Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Joseph D Barry, Harmanjit Singh Bansal, Oren Kraus, Mathias Wawer, Lassi Paavolainen, Markus D Herrmann, Mohammad Rohban, Jane Hung, Holger Hennig, John Con-cannon, Ian Smith, Paul A Clemons, Shantanu Singh, Paul Rees, Peter Horvath, Roger G Linington, and Anne E Carpenter. Data-analysis strategies for image-based cell profiling. *Nat. Methods*, 14(9):849–863, August 2017.
- Juan C Caicedo, Claire McQuin, Allen Goodman, Shantanu Singh, and Anne E Carpenter. Weakly supervised learning of feature embeddings for single cells in microscopy images. *IEEE CVPR*, 2018.
- Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *ICLR*, 2017.
- Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.
- Florian Fuchs, Gregoire Pau, Dominique Kranz, Oleg Sklyar, Christoph Budjan, Sandra Steinbrink, Thomas Horn, Angelika Pedal, Wolfgang Huber, and Michael Boutros. Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. *Mol. Syst. Biol.*, 6:370, June 2010.
- William J Godinez, Imtiaz Hossain, Stanley E Lazic, John W Davies, and Xian Zhang. A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics*, 33(13):2010–2019, July 2017.
- William J Godinez, Imtiaz Hossain, and Xian Zhang. Unsupervised phenotypic analysis of cellular images with multi-scale convolutional neural networks. *BioRxiv*, page 361410, 2018.

Peter Goldsborough, Nick Pawlowski, Juan C Caicedo, Shantanu Singh, and Anne Carpenter. Cytogan: Generative modeling of cell images. *bioRxiv*, page 227645, 2017.

Irina Higgins, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579*, 2016.

Thomas Horn, Thomas Sandmann, Bernd Fischer, Elin Axelsson, Wolfgang Huber, and Michael Boutros. Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. *Nat. Methods*, 8(4):341–346, March 2011.

Gregory R Johnson, Rory M Donovan-Maiye, and Mary M Maleckar. Generative modeling with conditional autoencoders: Building an integrated cell. *arXiv preprint arXiv:1705.00092*, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014.

Oren Z Kraus, Jimmy Lei Ba, and Brendan J Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, June 2016.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *ICML*, 2016.

Christina Laufer, Bernd Fischer, Maximilian Billmann, Wolfgang Huber, and Michael Boutros. Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nat. Methods*, 10(5):427–431, May 2013.

Christina Laufer, Bernd Fischer, Wolfgang Huber, and Michael Boutros. Measuring genetic interactions in human cells by RNAi and imaging. *Nat. Protoc.*, 9(10):2341–2353, October 2014.

Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nat. Methods*, 9(7):637, July 2012.

Vebjorn Ljosa, Peter D Caie, Rob Ter Horst, Katherine L Sokolnicki, Emma L Jenkins, Sandeep Daya, Mark E Roberts, Thouis R Jones, Shantanu Singh, Auguste Genovesio, Paul A Clemons, Neil O Carragher, and Anne E Carpenter. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J. Biomol. Screen.*, 18(10):1321–1329, December 2013.

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *ICLR*, 2016.

Claire McQuin, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, Beth A Cimini, Kyle W Karhohs, Minh Doan, Liya Ding, Susanne M Rafelski, Derek Thirstrup, and Others. CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.*, 16(7):e2005970, 2018.

Jessica L Ochoa, Walter M Bray, R Scott Lokey, and Roger G Linington. Phenotype-Guided natural products discovery using cytological profiling. *J. Nat. Prod.*, 78(9):2242–2248, September 2015.

Yoshikazu Ohya, Jun Sese, Masashi Yukawa, Fumi Sano, Yoichiro Nakatani, Taro L Saito, Ayaka Saka, Tomoyuki Fukuda, Satoru Ishihara, Satomi Oka, Genjiro Suzuki, Machika Watanabe, Aiko Hirata, Miwaka Ohtani, Hiroshi Sawai, Nicolas Fraysse, Jean-Paul Latgé, Jean M François, Markus Aebi, Seiji Tanaka, Sachiko Muramatsu, Hiroyuki Araki, Kintake Sonoike, Satoru Nogami, and Shinichi Morishita. High-dimensional and large-scale phenotyping of yeast mutants. *Proc. Natl. Acad. Sci. U. S. A.*, 102(52):19015–19020, December 2005.

Nick Pawlowski, Juan C Caicedo, Shantanu Singh, Anne E Carpenter, and Amos Storkey. Automating morphological profiling with generic deep convolutional networks. *BioRxiv*, page 085118, 2016.

Zachary E Perlman, Michael D Slack, Yan Feng, Timothy J Mitchison, Lani F Wu, and Steven J Altschuler. Multidimensional drug profiling by automated microscopy. *Science*, 306(5699):1194–1198, November 2004.

Benedikt Rauscher, Florian Heigwer, Luisa Henkel, Thomas Hielscher, Oksana Voloshanenko, and Michael Boutros. Toward an integrated map of genetic interactions in cancer cells. *Mol. Syst. Biol.*, 14(2):e7656, February 2018.

Felix Reisen, Amelie Sauty de Chalon, Martin Pfeifer, Xian Zhang, Daniela Gabriel, and Paul Selzer. Linking phenotypes and modes of action through High-Content screen fingerprints. *Assay Drug Dev. Technol.*, 13(7):415–427, September 2015.

Mohammad Hossein Rohban, Shantanu Singh, Xiaoyun Wu, Julia B Berhet, Mark-Anthony Bray, Yashaswi Shrestha, Xaralabos Varelas, Jesse S Boehm, and Anne E Carpenter. Systematic morphological profiling of human gene and allele function via cell painting. *Elife*, 6, March 2017.

Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017.

Xiongtao Ruan and Robert F Murphy. Evaluation of methods for generative modeling of cell and nuclear shape. *Bioinformatics*, December 2018.

S Singh, M-A Bray, T R Jones, and A E Carpenter. Pipeline for illumination correction of images for high-throughput microscopy. *J. Microsc.*, 256(3):231–236, December 2014.

Daniel W Young, Andreas Bender, Jonathan Hoyt, Elizabeth McWhinnie, Gung-Wei Chirn, Charles Y Tao, John A Tallarico, Mark Labow, Jeremy L Jenkins, Timothy J Mitchison, and Yan Feng. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.*, 4(1):59–68, January 2008.

# DavinciGAN: Unpaired Surgical Instrument Translation for Data Augmentation

**Kyungmoon Lee\***  
**Min-Kook Choi**  
**Heechul Jung**

KYUNGMOON@POSTECH.AC.KR  
MKCHOI@HUTOM.CO.KR  
HEECHUL@HUTOM.CO.KR

## Abstract

Recognizing surgical instruments in surgery videos is an essential process to describe surgeries, which can be used for surgery navigation and evaluation systems. In this paper, we argue that an imbalance problem is crucial when we train deep neural networks for recognizing surgical instruments using the training data collected from surgery videos since surgical instruments are not uniformly shown in a video. To address the problem, we use a generative adversarial network (GAN)-based approach to supplement insufficient training data. Using this approach, we could make training data have the balanced number of images for each class. However, conventional GANs such as CycleGAN and DiscoGAN, have a potential problem to be degraded in generating surgery images, and they are not effective to increase the accuracy of the surgical instrument recognition under our experimental settings. For this reason, we propose a novel GAN framework referred to as DavinciGAN, and we demonstrate that our method outperforms conventional GANs on the surgical instrument recognition task with generated training samples to complement the unbalanced distribution of human-labeled data.

**Keywords:** Generative adversarial network (GAN), image-to-image translation, self attention, data augmentation.

## 1. Introduction

To help surgeon’s decision making during the robotic surgery, providing surgical guidance like car navigation systems, based on the information extracted from the current surgery scene is necessary. Moreover, a surgery video should be analyzed to evaluate the robotic surgery after its operation. Recognizing surgical instruments is an essential process in such systems, and the information can be basically used for recognizing the current surgical phase (Twinanda et al., 2017).

In general, each surgical instrument is not used equally and uniformly in one operation. This leads to imbalance in terms of data collection, which is one of the critical problems in deep learning. In addition, since a certain tool is likely to show only in a specific environment, it can be said that context and background redundancy are high among the data for each tool. To address this issue, we propose an approach to translating an image. Generative adversarial network which is one of generative models is known for its ability to generate complex, high-dimensional data such as natural images (Goodfellow et al., 2014; Radford et al., 2016). With the advent of Conditional GAN, images can be created in the desired direction such as label-to-digit (Mirza and Osindero, 2014), text-to-image (Reed et al., 2016) and image-to-image (Isola et al., 2017; Choi et al., 2018).

---

\* Work done at Hutom

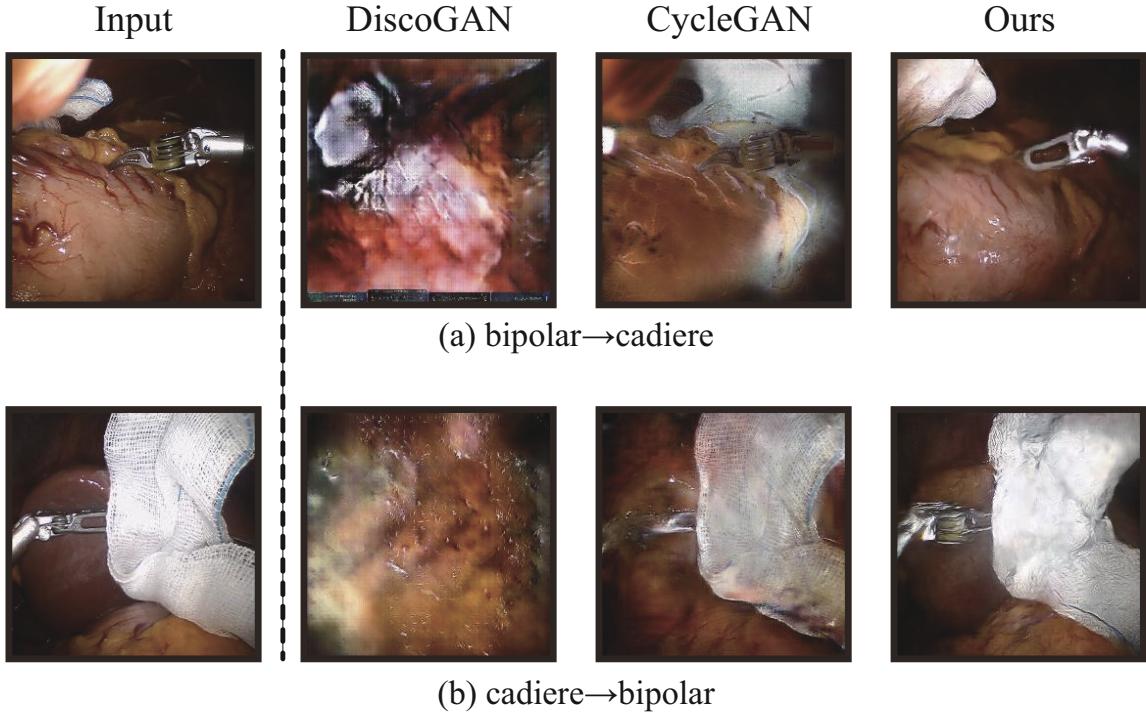


Figure 1: Results of the conventional works ([Kim et al., 2017](#); [Zhu et al., 2017](#)) and ours. Davinci-GAN gives the result with appearance changes and identical background simultaneously. (a) bipolar → cadiere, (b) cadiere → bipolar. From left to right: input, DiscoGAN ([Kim et al., 2017](#)), CycleGAN ([Zhu et al., 2017](#)) and DavinciGAN (ours).

In particular, unpaired image-to-image translation ([Zhu et al., 2017](#); [Kim et al., 2017](#)), which we address in this paper, has achieved impressive results recently.

However, these prior works failed easily when there are geometric changes between domains or the resolution of an input is high. Our goal is to address these issues as well as data imbalance problem between surgical instruments. To this end, our method, given an image, captures a candidate tool (e.g., cadiere) and transforms it into a target tool (e.g., bipolar), as shown in Figure 1. There are similar works ([Joo et al., 2018](#); [Tang et al., 2018](#)) that change gestures of a person while maintaining his/her identity, but they differ from ours in the sense that they deal with simple images without causing geometric changes.

Our main contributions are as follows:

1. We propose a new generative adversarial network, named as DavinciGAN that captures candidate daVinci instruments and transforms them into target daVinci instruments by making appearance changes.
2. We introduce background consistency loss using self-attention mechanism without ground truth mask data. With this loss, our network is encouraged to transform only the candidate tool to the target tool while maintaining background.

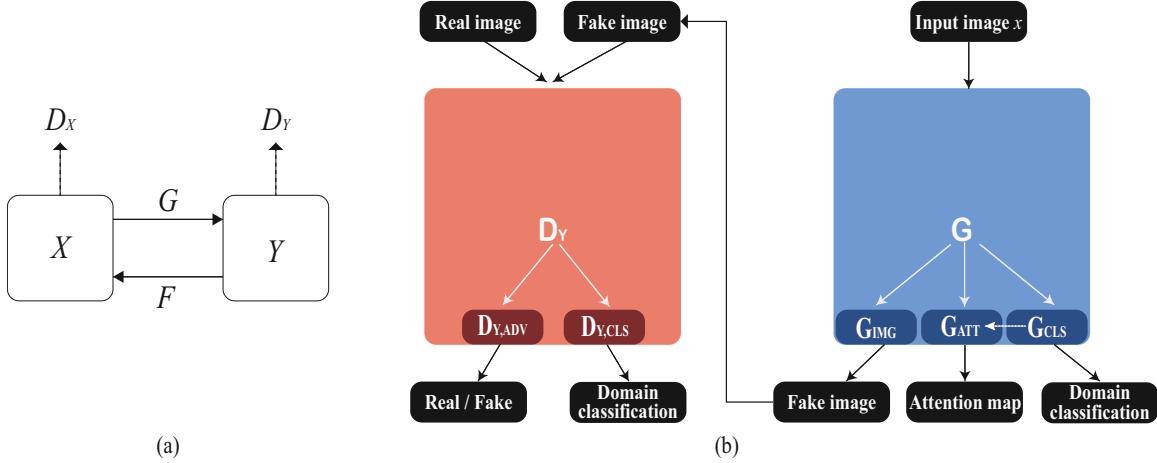


Figure 2: **Overall architecture of DavinciGAN.** (a) DavinciGAN consists of two generators  $G, F$  and two discriminators  $D_X, D_Y$ . (b) The figure on the left shows the discriminator  $D_Y$ , and the figure on the right shows the generator  $G$ . The  $F$  and  $D_X$  also follow this architecture.

3. We augment training data using GANs, and we show that the data augmentation using our DavinciGAN is the most effective to improving the instrument classification accuracies.
4. To the best of our knowledge, we first handle daVinci surgical instruments via image translation for data augmentation.

## 2. Methods

### 2.1. Architecture

Figure 2 illustrates the overall architecture of our DavinciGAN. Our goal is to find a mapping function to change the original surgical instrument in a source domain  $X$  to the desired surgical instrument in the target domain  $Y$  without background changes.

DavinciGAN consists of two generators  $G : x \rightarrow \{G_{IMG}(x), G_{ATT}(x), G_{CLS}(x)\}$  and  $F : y \rightarrow \{F_{IMG}(y), F_{ATT}(y), F_{CLS}(y)\}$ , where  $x \sim p_{data}(x)$  and  $y \sim p_{data}(y)$ .  $G_{IMG}$  represents a generator to generate fake image, and  $G_{ATT}$  produces an attention map computed from the predicted class score in domain classification and the feautre maps of  $G$  via weakly supervised learning technique (Zhou et al., 2016). Also,  $G_{CLS}$  is a classifier to classify domains. The  $F$  plays exactly the same role with  $G$ , but reverses two domains. Our DavinciGAN also has two discriminators  $D_X : x \rightarrow \{D_{X,ADV}(x), D_{X,CLS}(x)\}$  and  $D_Y : y \rightarrow \{D_{Y,ADV}(y), D_{Y,CLS}(y)\}$  and they also not only discriminate whether an input image is real or fake but also classify its domain.

### 2.2. Loss Function

We designed four loss functions such as adversarial loss, domain adversarial loss, background consistency loss and cycle consistency loss. The adversarial loss function is designed to make the

generated image distribution indistinguishable to the real data distribution. Also, the domain adversarial loss function is introduced with an auxiliary classifier to derive efficient topological changes. With the background consistency loss, our method can maintain the background information while changing the appearance of the tool. Lastly, the cycle consistency loss is helpful to reduce the space of possible mapping functions and guarantee one-to-one mapping between two domains.

### 2.2.1. ADVERSARIAL LOSS

We use an adversarial loss as a perceptual loss to enhance the naturalness of the generated images. By using this loss function, regardless of which output the auxiliary classifiers of discriminators give for domain classification, the generators learn to generate images indistinguishable from real images. In our work, we adopt LS-GAN loss ([Mao et al., 2017](#)) which is known to be advantageous for learning stability. We train  $G$  and  $F$  to maximize this objective and  $D_X$  and  $D_Y$  to minimize this objective.

$$\begin{aligned} L_{adv} = & \mathbb{E}_{x \sim p_{data}(x)} [\|1 - D_{X,ADV}(x)\|_2 + \mathbb{E}_{y \sim p_{data}(y)} [\|D_{X,ADV}(F_{IMG}(y))\|_2 \\ & + \mathbb{E}_{y \sim p_{data}(y)} [\|1 - D_{Y,ADV}(y)\|_2 + \mathbb{E}_{x \sim p_{data}(x)} [\|D_{Y,ADV}(G_{IMG}(x))\|_2]. \end{aligned} \quad (1)$$

### 2.2.2. DOMAIN ADVERSARIAL LOSS

Along with an adversarial loss, we introduce domain adversarial loss to lead to appearance changes. We added an auxiliary classifier for each discriminator to derive geometric changes by adversarial learning with the output of that auxiliary classifier.

$$L_{D,CLS} = BCE(D_X) + BCE(D_Y), \quad (2)$$

where

$$BCE(N) \triangleq \mathbb{E}_{y \sim p_{data}(y)} [-\log(N_{CLS}(y))] + \mathbb{E}_{x \sim p_{data}(x)} [-\log(1 - N_{CLS}(x))]. \quad (3)$$

As a two-class classification problem, we set the label of domain  $X$  to 0 and the label of domain  $Y$  to 1. We train discriminators to classify not only accurate domains of real data, but also real  $x$  closer to zero than fake  $x'$  generated from  $F$  and real  $y$  closer to one than fake  $y'$  generated from  $G$ . On the other hand,  $G$  and  $F$  try to fool discriminators to classify  $y'$  closer to one than  $y$  and  $x'$  closer to zero than  $x$ , respectively.  $D_X$  and  $D_Y$  try to minimize Eq (2) and Eq (4) and  $G$  and  $F$  try to maximize Eq (4).

$$\begin{aligned} L_{CLS-ADV} = & \mathbb{E}_{x \sim p_{data}(x), y \sim p_{data}(y)} [D_{X,CLS}(x) - D_{X,CLS}(F_{IMG}(y)) + \\ & D_{Y,CLS}(G_{IMG}(x)) - D_{Y,CLS}(y)], \end{aligned} \quad (4)$$

### 2.2.3. BACKGROUND CONSISTENCY LOSS

$$L_{G,CLS} = BCE(G) + BCE(F). \quad (5)$$

Unlike traditional image translation tasks, we want to cross-domain via transforming only the instrument while maintaining the background. In our method, as in the case of the discriminator, we added an auxiliary classifier for each generator to find out which region is important to classify the domain via weakly supervised learning technique ([Zhou et al., 2016; Singh and Lee, 2017](#)) by minimizing Eq (5). We set the label of domain  $X$  to 0 and the label of domain  $Y$  to 1 exactly like 2.2.2. By utilizing that region as a format of attention map, we try to find out where the

instrument is in the given image without ground truth mask and transform the target instance only, while maintaining the background by minimizing Eq (6).

$$L_{BG-CONSIST} = \mathbb{E}_{x \sim p_{data}(x), y \sim p_{data}(y)} [\| (x - G_{IMG}(x)) \otimes (1 - G_{ATT}(x)) \|_1 + \| (y - F_{IMG}(y)) \otimes (1 - F_{ATT}(y)) \|_1]. \quad (6)$$

#### 2.2.4. CYCLE CONSISTENCY LOSS

In the task of unpaired image-to-image translation, it is known that adversarial losses are not enough to guarantee the mapping between an input and the desired output because random data of target domain distribution can be generated when only adversarial loss functions are optimized. Since the task we address is to change only the instrument while maintaining the background, an input  $i$  and  $G_{IMG}(i)$  or  $F_{IMG}(i)$  must have one-to-one correspondence in theory. Cycle consistency loss is a great help in this context and we train  $G$  and  $F$  to minimize this objective.

$$L_{CYC-CONSIST} = \mathbb{E}_{x \sim p_{data}(x), y \sim p_{data}(y)} [\| x - F_{IMG}(G_{IMG}(x)) \|_1 + \| y - G_{IMG}(F_{IMG}(y)) \|_1]. \quad (7)$$

#### 2.2.5. FULL OBJECTIVE

To sum it up, the full objective functions to optimize discriminators and generators are as follows, respectively.

$$L_D = L_{adv} + \lambda_{CLS} * (L_{D,CLS} + L_{CLS-ADV}). \quad (8)$$

$$L_G = -L_{adv} + \lambda_{CLS} (L_{G,CLS} - L_{CLS-ADV}) + \lambda_{BG} L_{BG-CONSIST} + \lambda_{CYC} L_{CYC-CONSIST}. \quad (9)$$

For hyper-parameter setting, we set  $\lambda_{CLS} = 1$ ,  $\lambda_{BG} = 10$  and  $\lambda_{CYC} = 10$ .

### 3. Experiments and results

#### 3.1. Dataset

With 8 surgery videos using the daVinci Surgical System, we label frames having only one corresponding instrument. Since we found that bipolar is common and cadiere is rare relatively among all surgical instruments, we chose these two instruments to conduct experiments and finally, we built our surgical instrument dataset, consisting of 29,207 images where 15,344 are bipolar and 13,863 are cadiere.

#### 3.2. Experimental settings

We compare DavinciGAN with baseline models such as DiscoGAN and CycleGAN. The size of input and output images in our experiment is  $256 \times 256$ . For this setting, we added extra parameters to DiscoGAN addressing  $64 \times 64$  size originally to equalize the number of learnable parameters. We adopt an Encoder-Decoder architecture which can be better in realizing appearance changes for our generators. Our discriminator uses  $32 \times 32$  PatchGAN (Isola et al., 2017; Li and Wand, 2016; Ledig et al., 2017; Zhu et al., 2017) to classify whether patches are real or fake with the spatial

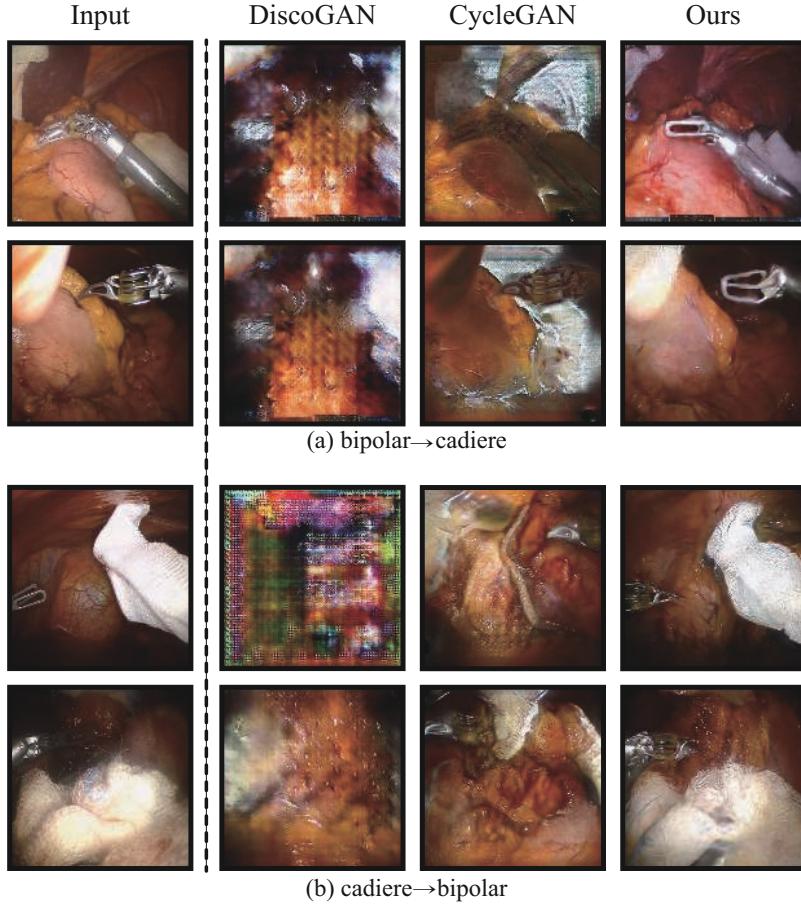


Figure 3: **Additional translation results** (a): bipolar → cadiere and (b): cadiere → bipolar. From left to right: input, DiscoGAN (Kim et al., 2017), CycleGAN (Zhu et al., 2017), DavinciGAN (ours).

information. Our network also used CBAM bottlenecks (Woo et al., 2018) between layers to get better performance of self attention.

We have chosen a mini-batch size of 8, and only horizontal flip was used as a data augmentation technique. Furthermore, we use the Adam optimizer (Kingma and Adam, 2015) with learning rate of 0.0002,  $\beta_1$  of 0.5 and  $\beta_2$  of 0.999. All models are implemented using Tensorflow and trained on a NVIDIA TITAN Xp GPU.

### 3.3. Results

#### 3.3.1. QUALITATIVE RESULTS

Figure 3 shows our qualitative results on our surgical instrument dataset. DavinciGAN generated more convincing results than two baselines, DiscoGAN and CycleGAN. While maintaining the overall structure of background and surgical aids such as gauze, DavinciGAN transformed candidate

instruments into target instruments by causing shape changes. In spite of trying various settings such as tuning the learning rate and the learning ratio between generator and discriminator per iteration, we found that DiscoGAN failed into mode collapse. This seems to be due to the fact that there is no constraint on the first translation such as background consistency loss function while handling complex images with organs, blood vessels and surgical aids (e.g., gauze, needle). CycleGAN adopts a fully convolutional ResNet structure (Johnson et al., 2016) which is known to be advantageous in generating high resolution images with little change in input. Although it maintains the structure of an input image well, it has difficulties in making variations to the candidate instrument with no constraint such as domain adversarial loss function.

### 3.3.2. QUANTITATIVE RESULTS

Table 1 shows the quantitative results of instrument classification utilizing the fixed real data and synthetic data generated by each model as training data. In consideration of the difficulty to collect surgical image data, the experiment was conducted with limited training data. We trained GAN-based models on 3,987 images where 2,419 are bipolar and 1,568 are cadiere from three videos. For instrument classification task, we used 500 real data and 500 synthetic data for each instrument as training data while leaving all images of the remainder five videos as test data. For all cases, ResNet50 (He et al., 2016) trained until convergence. As a result, although DavinciGAN used less parameters than two baselines, it was superior to baselines in test accuracy and even competitive with additional real data.

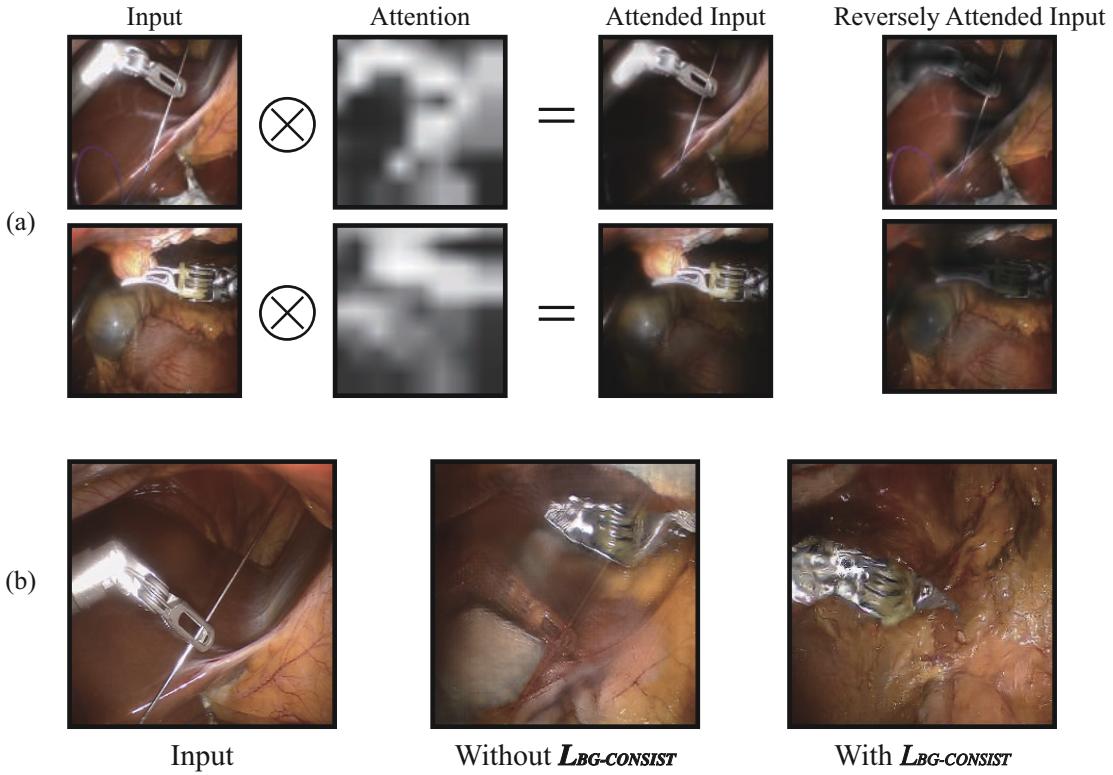
Table 1: Classification performances and the number of parameters for each method.

Dataset	Method	# of parameters	Accuracy (%)
Real 1000	-	-	58.84
Real 1000 + Synthetic 1000	DiscoGAN	67M	57.91
Real 1000 + Synthetic 1000	CycleGAN	56M	58.61
Real 1000 + Synthetic 1000	DavinciGAN	31M	61.34
Real 2000	-	-	62.31

## 4. Discussion

### 4.1. Self attention via weakly supervised learning

Figure 4 (a) visualizes attention maps from generators and reversely attended input for each instrument. Reversely attended inputs which are utilized for the background consistency loss tend to hide instruments’ head which is discriminative features to identify which instrument it is. We introduced self attention mechanism via weakly supervised learning to capture the position of instrument to transform without ground truth mask data. However, there are failure cases caused by the failure of attention. When generators classify the instruments, it tends to attend to the surgical aids, such as a gauze or the User Interface of daVinci surgical system, as well as the instrument’s shape. As a result, we found that undesired attention maps are generated as shown in Figure 5. Since certain instruments tend to appear only in certain situations, classifier also tends to predict an output using other discriminative features rather than the instrument itself. As mentioned in section 3.3.2,

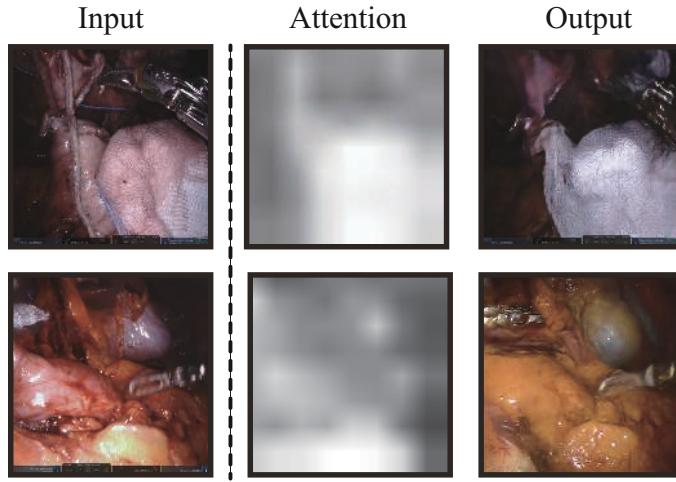


**Figure 4: How the generated attention is utilized in DavinciGAN and the visual analysis of the background consistency loss.** (a): Attention maps produced by generators. Attended input shows discriminative features such as the head of instruments. On the other hand, reversely attended input which is utilized for the background consistency loss tends to hide discriminative features of the instrument. (b): Without the background consistency loss, the bipolar is generated at an arbitrary position, but with background consistency loss, it is generated at the position of cadiere.

limited data was used as training data due to the consideration of challenges to collect rich surgical video data. Surgical videos of more diverse surgeons will increase the appearance of instruments in orthogonal contexts, which can improve the performance of attention.

#### 4.2. The effectiveness of background consistency loss

Interestingly, as shown in Figure 4 (b), the background consistency loss shows an intuitive result. Without the background consistency loss, our network generated a synthetic bipolar in the upper right side while hiding the cadiere by changing its color. This is because generators can fool discriminators by generating target instruments at any location, which is not a desired output for us. However, with the background consistency loss, our network shows its capability to generate the target instrument at the desired position.



**Figure 5: Failure cases of DavinciGAN with undesired attention maps.** Attention maps from generators focus on gauze and User Interface of daVinci Surgical System, not on the shapes of instruments. With these undesired attention maps, generators produced undesired outputs.

## 5. Conclusion

In this paper, we propose a novel generative adversarial network, DavinciGAN which transforms only surgical instrument while maintaining the background. This is achieved by the domain adversarial loss function and the background consistency loss function. We have showed qualitative and quantitative results on how useful generated data can be as training data compared to two baselines such as DiscoGAN and CycleGAN. One of advantages with our method is that DavinciGAN utilizes the self attention mechanism through weakly supervised learning approach so that we do not require any other annotation data like segmentation masks.

## References

- Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- Donggyu Joo, Doyeon Kim, and Junmo Kim. Generating a fusion image: One's identity and another's shape. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.
- D Kinga and J Ba Adam. A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.
- Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.
- Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.
- Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

Hao Tang, Wei Wang, Dan Xu, Yan Yan, and Nicu Sebe. Gesturegan for hand gesture-to-gesture translation in the wild. *arXiv preprint arXiv:1808.04859*, 2018.

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 2017.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2018.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

# Dense Segmentation in Selected Dimensions: Application to Retinal Optical Coherence Tomography

**Bart Liefers<sup>1,2</sup>**

BART.LIEFERS@RABDOUDUMC.NL

**Cristina González-Gonzalo<sup>1,2</sup>**

**Caroline Klaver<sup>3,4</sup>**

**Bram van Ginneken<sup>2</sup>**

**Clara I. Sánchez<sup>1,2,3</sup>**

<sup>1</sup> A-Eye Research Group, Radboudumc, Nijmegen, The Netherlands

<sup>2</sup> Diagnostic Image Analysis Group, Radboudumc, Nijmegen, The Netherlands

<sup>3</sup> Department of Ophthalmology, Radboudumc, Nijmegen, The Netherlands

<sup>4</sup> Ophthalmology & Epidemiology, Erasmus MC, Rotterdam, The Netherlands

## Abstract

We present a novel convolutional neural network architecture designed for dense segmentation in a subset of the dimensions of the input data. The architecture takes an  $N$ -dimensional image as input, and produces a label for every pixel in  $M$  output dimensions, where  $0 < M < N$ . Large context is incorporated by an encoder-decoder structure, while funneling shortcut subnetworks provide precise localization. We demonstrate applicability of the architecture on two problems in retinal optical coherence tomography: segmentation of geographic atrophy and segmentation of retinal layers. Performance is compared against two baseline methods, that leave out either the encoder-decoder structure or the shortcut subnetworks. For segmentation of geographic atrophy, an average Dice score of  $0.49 \pm 0.21$  was obtained, compared to  $0.46 \pm 0.22$  and  $0.28 \pm 0.19$  for the baseline methods, respectively. For the layer-segmentation task, the proposed architecture achieved a mean absolute error of  $1.305 \pm 0.547$  pixels compared to  $1.967 \pm 0.841$  and  $2.166 \pm 0.886$  for the baseline methods.

**Keywords:** Segmentation, Retina, OCT

## 1. Introduction

Many applications of deep convolutional neural networks (CNNs) in medical imaging can be formulated as either a classification or a segmentation problem (Litjens et al., 2017). Classification problems can be defined as a mapping from an input of  $N$  spatial dimensions (usually  $N$  is 2 or 3, for 2D or 3D images) to an output without any spatial dimension, while for segmentation problems the  $N$ -dimensional input is mapped to an  $N$ -dimensional output. However, some medical applications require the projection of  $N$ -dimensional images to  $M$ -dimensional output, where  $0 < M < N$ . Problems that can be formulated as extracting a 2D manifold from a 3D volume include the detection of boundary surfaces or fissures, such as separating the cerebral hemispheres in MRI (Liang et al., 2007), segmenting pulmonary fissures in CT (Wang et al., 2006), or delineating the diaphragm in CT (Rangayyan et al., 2008). Similarly, a 1D line can be extracted from a 2D image (e.g. retinal layers (Fang et al., 2017)). Direct application of well-known classification or segmentation network architectures to this class of problems comes with limitations. Classification networks do not main-

tain any spatial information in their output values, so adapting them to produce a label for each pixel in the desired output dimensions is not feasible in practice. Segmentation networks, on the other hand, produce a dense  $N$ -dimensional output for all pixels, but there is no natural way to enforce output in just  $M$ -dimensions. This is problematic in case we do not have a label for all pixels.

In this paper we present a novel CNN architecture that produces dense predictions only in the desired output dimensions, maintaining spatial correlation. To the best of our knowledge, this is the first CNN architecture that is specifically designed for this task. Large context from  $N$  input dimensions is incorporated by an encoding network. A selective decoding is applied only in the required  $M$  output dimensions.

The encoder-decoder paradigm has been applied successfully to many segmentation problems in medical imaging, where U-Net (Ronneberger et al., 2015) (or a 3D variation (Çiçek et al., 2016)) is especially popular. In this network several layers of downsampling are applied to the input image in a contracting or encoding path. Subsequently, the original resolution is reconstructed in an expanding or decoding path that is connected via shortcut connections to the feature maps of the corresponding resolution in the contracting path. These shortcut connections are important for precise localization (Drozdzal et al., 2016).

In our network architecture, direct shortcut connections between the encoding and decoding blocks are not possible, because of a mismatch in the dimensions of the feature maps. This issue is resolved by adding funneling shortcut subnetworks between the encoder and decoder that contract only in the dimensions that are spliced out.

Although the proposed general formulation of the network architecture is applicable to any  $N$  and  $M$  ( $0 < M < N$ ), we will focus in this paper on the case where  $N = 2$  and  $M = 1$ . More specifically, we focus on application to retinal optical coherence tomography (OCT), where we predict a label for every vertical column (A-scan) in each slice of the OCT volume (B-scan). We demonstrate its applicability to two problems: segmentation of retinal layers, and segmentation of geographic atrophy (GA), and compare performance against two baseline models.

## 2. Background

The retina is a layered structure in the back of the eye that converts light into a neural response that provides vision. This layered structure can be visualized non-invasively using OCT, an imaging technique based on low coherence interferometry. OCT is commonly used to generate stacks of cross-sectional 2D images (B-scans), where every column in the image (an A-scan) represents a single acquisition entity. Often, the distance between slices is much larger than the distance between A-scans, so the acquired volume is highly anisotropic.

The proposed network architecture is particularly suitable for application to OCT, due to the nature of the image acquisition and the horizontally-layered structure of the retina. The pixel intensities within A-scans are strongly correlated, and the presence or absence of certain features can be identified for each of them, without specifying their location within the A-scan. In order to accurately classify each A-scan, contextual information from surrounding A-scans is usually required. Some examples in this category include segmentation of atrophic areas (Hu et al., 2013; Niu et al., 2016), localization of anatomical landmarks such as the fovea (Liefers et al., 2017) or segmentation of vessels (Tan et al., 2018). Another application typical for retinal OCT is the segmentation of retinal layers (Fang et al., 2017; Kugelman et al., 2018). These layers are horizontally stacked, so the boundary between them can be encoded as the vertical pixel index of the transition between two

layers. Hence, these problems can all be formulated as taking a 2D image as input and generating a label or a regression output for each column in the image, yielding a 1D output.

The two applications we focus on in this paper are segmentation of GA and retinal layers. GA occurs as an end stage in age-related macular degeneration (AMD) and is characterized by loss of retinal tissue and pigment. Next to absence of certain layers, this can be identified on OCT as a hyper-reflective area below the retina (referred to as hypertransmission). Segmentation of the individual layers of the retina allows to measure their thickness. Both applications can help ophthalmologists to diagnose disease status and decide treatment options.

### 3. Method

#### 3.1. Data

Data for the application of the proposed model to segmentation of GA was collected from the Rotterdam Study ([Hofman et al., 2007](#)). This data set contains 55 OCT volumes acquired with a Topcon system, each containing 128 B-scans of either  $512 \times 885$  pixels or  $512 \times 650$  pixels (width  $\times$  height). Manual annotations were made by four to five experienced graders in consensus using a multimodal annotation workstation. The graders did not directly annotate the OCT volume, but delineated the area of GA on 2D en-face retinal images (color fundus, and for some cases also autofluorescence or infrared). Image registration of these en-face images to the OCT volume was used to efficiently obtain a label (presence or absence of GA) for every A-scan in the volume. The data was split in a training set of 25 volumes, a validation set of 10 volumes, and a test set of 20 volumes.

For the layer-segmentation problem we used the publicly-available data set ([Farsiu et al., 2014](#)), containing annotations for three retinal layers: inner limiting membrane (ILM), retinal pigment epithelium drusen complex (RPEDC) and Bruch's membrane (BM). This data set contains OCT volumes for 269 AMD patients and 115 normal subjects. It was split in a training set (159 AMD, 5 normal), a validation set (10 AMD, 10 normal) and a test set (100 AMD, 100 normal). OCT volumes in this data set were acquired using a Bioptigen system, and contain 100 B-scans of  $1000 \times 512$  pixels each.

#### 3.2. Network architecture

The proposed network architecture follows an encoder-decoder structure, providing a large contextual window, with funneling subnetworks that provide local context. In this section we will focus on the case where the model takes a 2-dimensional input, and produces a label or a regression output for every vertical column in the image. This specific instance of the proposed generic network architecture is visualized in Figure 1.

The proposed architecture applies 9 levels of downsampling, which effectively reduces the input image from  $512 \times 512$  pixels to  $1 \times 1$  pixel in the deepest layer. This allows the model to capture information from a large contextual window. Residual connections are used to facilitate effective training of the deep network architecture ([He et al., 2015](#); [De Fauw et al., 2018](#)).

Next to a path with  $2 \times 2$  downsampling operations, separate downsampling paths are added to the encoder path at every resolution. In these downsampling paths  $2 \times 1$  (height  $\times$  width) operations are used, therefore only reducing the vertically resolution. These paths constitute the funneling

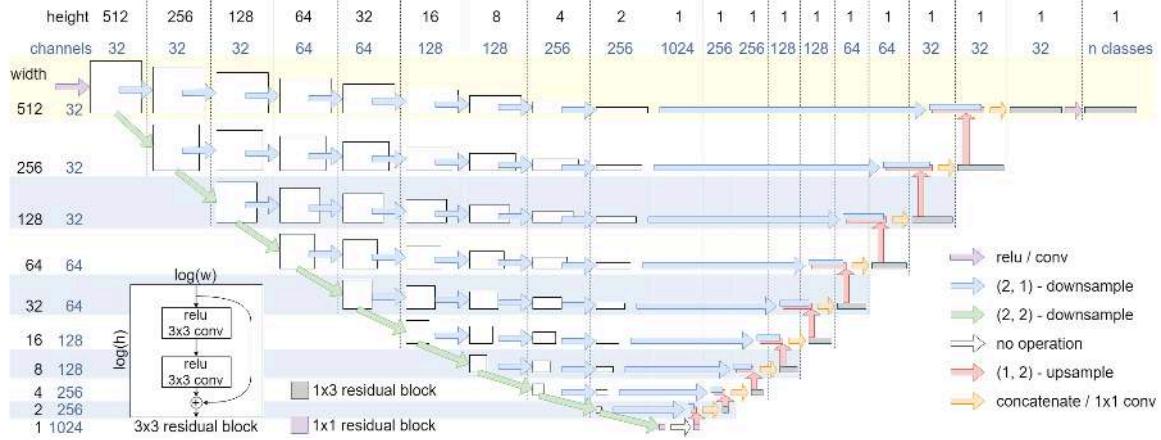


Figure 1: Proposed network architecture for 2D images to 1D segmentation. The white squares represent the size (log-scale) of the feature maps. Within each square, two convolutions are applied in a residual manner. For visualization purposes, the feature maps with width/height 1, have the same size as the feature maps with size 2. In the upsampling path, the residual blocks have height 1, and the convolutional kernels are of size  $1 \times 3$ .

shortcut subnetworks that should provide the model with better localization abilities. Downsampling is performed by either  $2 \times 2$  convolutions with stride  $2 \times 2$  or  $2 \times 1$  convolutions with stride  $2 \times 1$ .

At the bottom of the network, 2 residual blocks with  $1 \times 1$  convolutions are applied. The upsampling path consists of  $1 \times 2$  upsampling operations, followed by  $1 \times 2$  convolutions with stride  $1 \times 2$ . Before entering the first residual block, a  $3 \times 3$  convolution with 32 filters is applied to match the required number of features for the summation operation in the first residual block. After concatenating the feature maps of the upsampling path and the shortcut networks, a  $1 \times 1$  convolution is applied to match the required number of features for the next residual block. The kernel size of the convolutions in the downsampling path is  $3 \times 3$ ; at the bottom of the network,  $1 \times 1$ ; and in the upsampling path,  $1 \times 3$ .

Although the height of the input images is fixed by the network architecture, their width can be any multiple of 512. This is exploited in the application to layer segmentation, where we apply the same architecture to images that are 1024 pixels wide.

### 3.3. Experimental design

We compare our proposed network architecture to two baseline models, that are constructed by leaving out either the encoder-decoder structure, or the shortcut subnetworks.

The first baseline model, referred to as *base 1*, applies downsampling only in the dimensions that will be spliced out, until the feature maps are sufficiently compressed and can be passed to a classification layer. This architecture corresponds to part of the proposed architecture highlighted in the top yellow row in Figure 1. For this model, before the output layer, we also added a classification

part that consists of two  $1 \times 1$  residual blocks with 1024 filters each, in order to mimic classification networks more closely.

The second baseline model, referred to as *base 2*, leaves out the shortcut subnetworks. This architecture can be created by following the green and red arrows in Figure 1, ignoring the blue and orange arrows.

### 3.4. Training procedure

#### 3.4.1. GA SEGMENTATION

Images were cropped to  $512 \times 512$  pixels. The vertical position was centered on the row in the original image with the highest cumulative pixel intensity, to ensure the retina was visible in the output image. The images were augmented using random horizontal and vertical translations and horizontal flipping. Additionally, pixel intensities were modified using gamma corrections. A sigmoid function was applied to the output image of  $512 \times 1$  pixels, to obtain a normalized binary label for every A-scan. Mean log loss (binary cross-entropy) was used a loss function. The models were trained for  $2 \times 10^4$  iterations using the Adam optimizer, with a learning rate of  $1 \times 10^{-5}$ , divided by 2 after  $10^4$  iterations, on batches of 4 images.

#### 3.4.2. LAYER SEGMENTATION

Input images were padded to  $1024 \times 512$  pixels. Random horizontal and vertical translation, and random horizontal flips were used as data augmentation. The output layer has three channels, each producing an image of size  $1024 \times 1$ , representing the normalized pixel index of the three layers (ILM, RPEDC, BM). An output of 0 for a specific layer in an A-scan is interpreted as the layer being completely at the top of the image (pixel 0), while an output of 1 is interpreted as the layer being completely at the bottom (pixel 512). No non-linearity was applied to the output layer.

Mean squared error was used as loss function. However, to calculate the loss for a given layer, we only took into account those A-scans where the squared error was larger than the mean squared error of all A-scans for that layer. This modification was made because the contribution to the loss of many small errors may dominate some of the larger, more localized errors. Additionally, images were more frequently sampled for training if their loss was larger than the median loss of the last 100 iterations. The model was trained for  $4 \times 10^5$  iterations, one image per batch, using the Adam optimizer and a learning rate of  $2 \times 10^{-5}$ , divided by 2 every  $10^5$  iterations.

## 4. Results

On the GA data set we calculated a single Dice score per OCT-volume. The proposed model obtained a mean ( $\pm$  std) score of  $0.49 \pm 0.21$ , compared to  $0.46 \pm 0.22$  for base 1 and  $0.28 \pm 0.19$  for base 2. The proposed model performed significantly better than the two baseline models ( $p < 0.01$ , paired t-test). Examples of the obtained GA segmentations per B-scan can be found in Figure 2. A selected set of predictions for full volumes can be found in Figure 3.

On the layer-segmentation task, the proposed model obtained a mean absolute difference to the reference standard (averaged over all layers) of  $1.305 \pm 0.547$  pixels, compared to  $1.967 \pm 0.841$  and  $2.166 \pm 0.886$  for base 1 and base 2, respectively. Table 1 summarizes the performance of the models in more detail. An example with predictions for the different models on a single B-scan can be found in Figure 4.

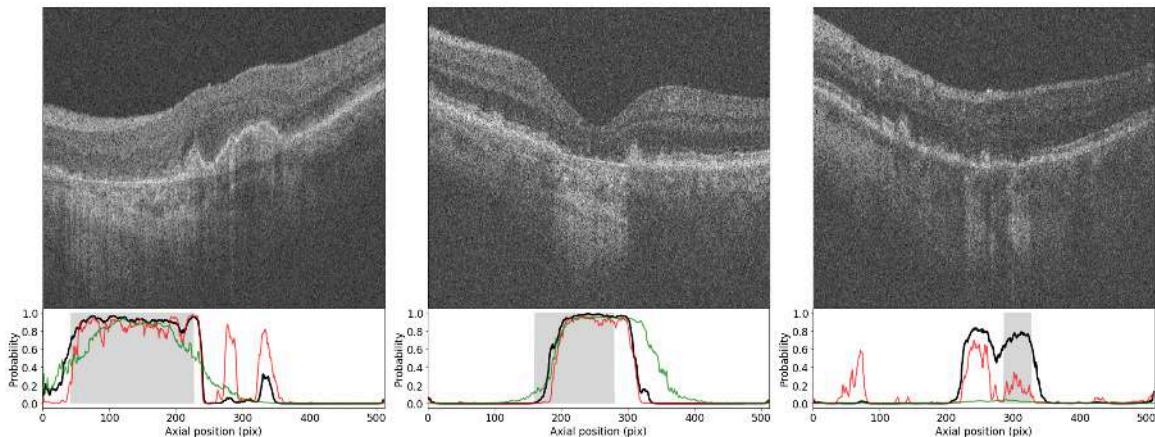


Figure 2: Predictions for GA for 3 different B-scans. The graphs show the reference standard (gray area), and predicted GA probabilities. The black line represents the proposed model; the red line, base 1; and the green line, base 2.

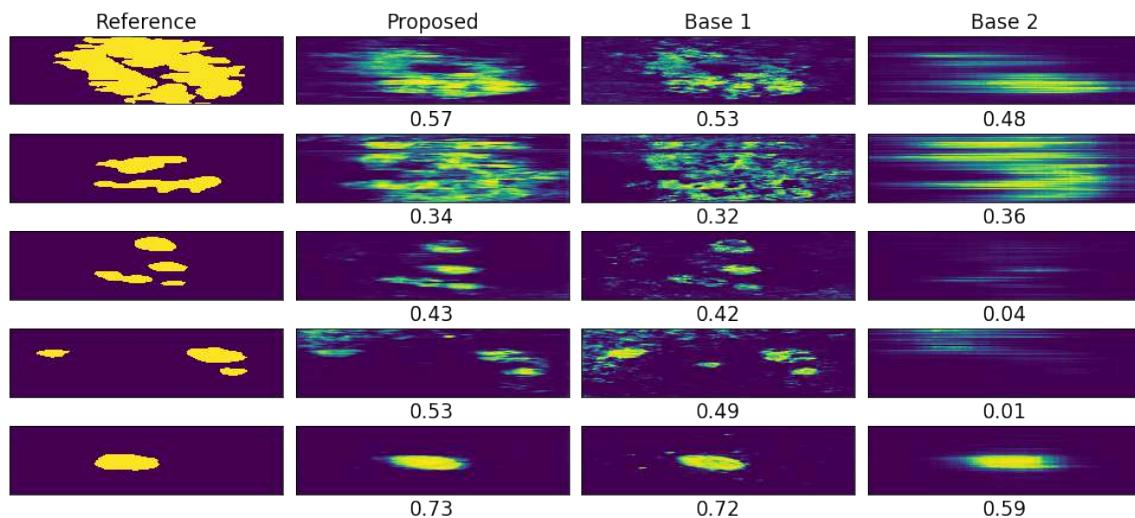


Figure 3: Selection of predictions of GA area for full OCT-volumes. Each horizontal line in the images represents a B-scan. The left column represents the reference standard, the heatmaps in the second to fourth column represent the probabilities generated by the proposed model, base 1, and base 2, respectively. Numbers below the images are the Dice scores.

Table 1: Results on the layer-segmentation task. Values represent mean absolute difference between prediction and annotation in pixels.

	Control			AMD			Overall
	ILM	RPEDC	BM	ILM	RPEDC	BM	
Proposed	0.840	1.280	1.227	1.055	1.568	1.858	1.305
Base 1	1.031	1.507	2.246	1.200	2.267	3.553	1.967
Base 2	1.880	1.762	1.673	2.203	2.785	2.693	2.166

## 5. Discussion

We proposed a novel CNN architecture for dense segmentation in selected output dimensions. The proposed architecture was validated on two applications in retinal OCT, where it achieved superior performance compared to two baseline models on both tasks. We attribute this to the ability of the proposed architecture to simultaneously make use of a large contextual window, and local context through funneling shortcut connections. This was demonstrated on the layer-segmentation task, where base 1 performed worse especially in localization of BM in AMD, for which a large contextual window is required in presence of abnormalities. Base model 2 can make use of a large context, but appears to be less accurate in general. This may be due to missing shortcut connections, which hampers its ability to perform exact localization. This observation is illustrated by the example in Figure 4. For the GA segmentation we observe similar results. For example, in the left image of Figure 2, the misclassifications in the center of the image of base 1 may be due to lack of context, while base 2 is unable to accurately delineate the borders of the atrophic region.

The Dice scores for the GA segmentation task are relatively low. This is partly due to the inherent difficulty of the task, which may have led to inaccuracies in the reference standard. Moreover, the actual GA area sometimes does not align perfectly with the GA that is visible in the B-scan due to registration errors. This is demonstrated in Figure 2 in the center and right images, where the reference grading does not align well with the observed hypertransmissive region. Visually judging the predictions of the proposed model in Figure 3, however, seems to indicate that even with relatively low Dice scores, the predicted GA areas are plausible.

A limitation of this work is that the proposed general architecture for problems that require mapping an  $N$ -dimensional input to an  $M$ -dimensional output, has only been validated for the 2D to 1D mapping of retinal OCT data. In future work, we hope to apply the proposed architecture also for the 3D to 2D case. Although the anisotropic resolution of OCT volumes is a hurdle in direct application of 3D convolutions, we do see potential applications of the proposed architecture for isotropic images in OCT angiography. Here it could be used, for example, in the generation of 2D images representing perfusion density in different vascular plexi from 3D volumetric data.

## 6. Conclusion

We demonstrated the applicability of a new CNN architecture to two problems in retinal OCT, where its value was warranted by the unique properties of the retina (its horizontally-layered structure), and the image acquisition of OCT (where each A-scan represents a column of information on a retinal

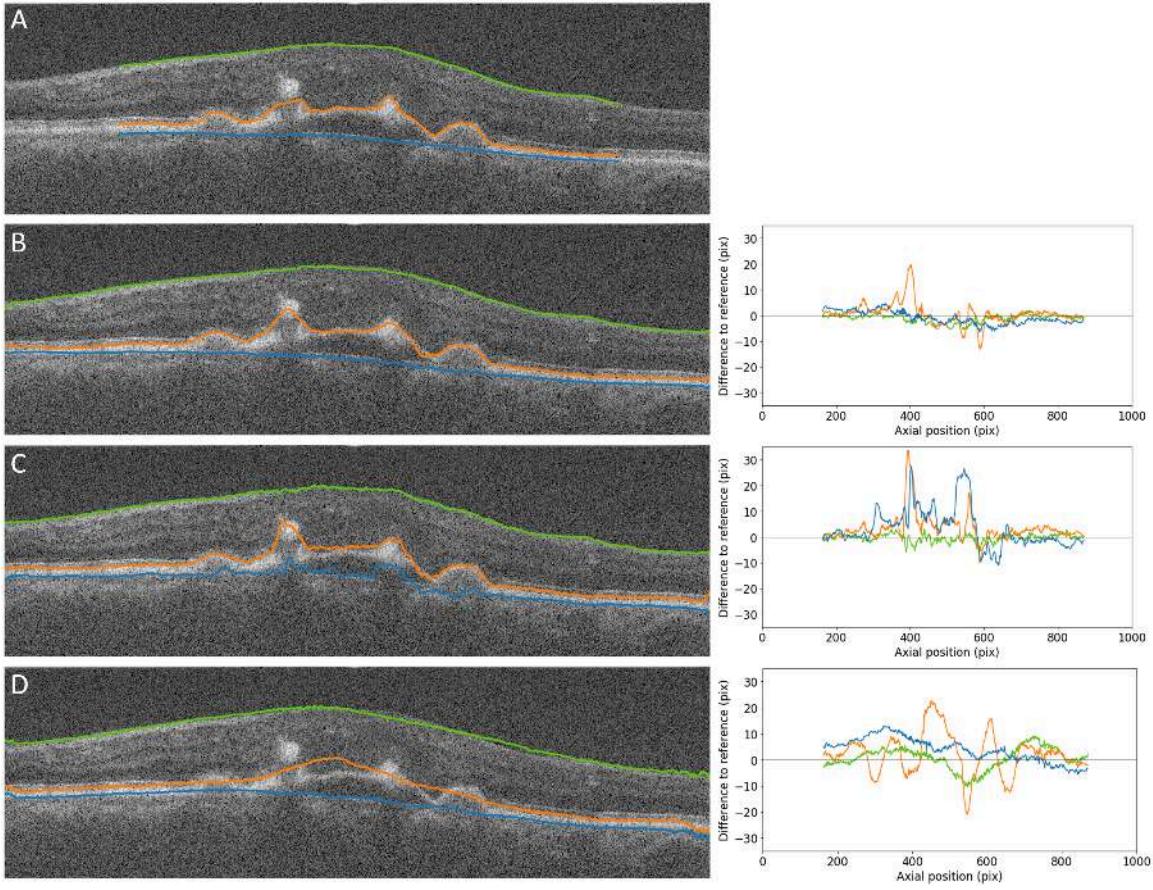


Figure 4: Example result for the layer-segmentation task. Image A shows the reference standard, which is only provided for the central part of the image. Images B, C and D represent the proposed model, base 1, and base 2, respectively. The graphs on the right depict the difference in pixels with the reference for this example for the ILM, RPEDC and BM, in green, orange and blue, respectively.

surface). The proposed model consistently outperformed two baseline models, which indicates that the combination of a large contextual window, provided by an encoder-decoder structure, and accurate localization, provided by funneling shortcut subnetworks, is beneficial.

## References

- Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.
- J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O’Donoghue, D. Visentin, et al. Clinically applicable deep learning for diagnosis

- and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018.
- M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer, 2016.
- L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomedical Optics Express*, 8(5):2732–2744, 2017.
- S. Farsiu, S. J. Chiu, R. V. O’Connell, F. A. Folgar, E. Yuan, J. A. Izatt, and C. A. Toth. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology*, 121(1):162–172, 2014.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015.
- A. Hofman, M. M. B. Breteler, C. M. van Duijn, G. P. Krestin, H. A. Pols, B. H. C. Stricker, Henning T., A. G. Uitterlinden, J. R. Vingerling, and J. C. M. Witteman. The Rotterdam Study: objectives and design update. *European Journal of Epidemiology*, 22(11):819–829, 2007.
- Z. Hu, G. G. Medioni, M. Hernandez, A. Hariri, X. Wu, and S. R. Sadda. Segmentation of the geographic atrophy in spectral-domain optical coherence tomography and fundus autofluorescence images. *Investigative Ophthalmology & Visual Science*, 54(13):8375, 2013.
- J. Kugelman, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins. Automatic segmentation of OCT retinal boundaries using recurrent neural networks and graph search. *Biomedical Optics Express*, 9(11):5759–5777, 2018.
- L. Liang, K. Rehm, R. P. Woods, and D. A. Rottenberg. Automatic segmentation of left and right cerebral hemispheres from mri brain volumes using the graph cuts algorithm. *NeuroImage*, 34 (3):1160–1170, 2007.
- B. Liefers, F. G. Venhuizen, V. Schreur, B. van Ginneken, C. Hoyng, S. Fauser, T. Theelen, and C. I. Sánchez. Automatic detection of the foveal center in optical coherence tomography. *Biomedical Optics Express*, 8(11):5160–5178, 2017.
- G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, and C.I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.
- S. Niu, L. de Sisternes, Q. Chen, T. Leng, and D. L. Rubin. Automated geographic atrophy segmentation for SD-OCT images using region-based cv model via local similarity factor. *Biomedical Optics Express*, 7(2):581–600, 2016.
- R. M. Rangayyan, R. H. Vu, and G. S. Boag. Automatic delineation of the diaphragm in computed tomographic images. *Journal of Digital Imaging*, 21(1):134–147, 2008.
- O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241, 2015.

- B. Tan, A. Wong, and K. Bizheva. Enhancement of morphological and vascular features in OCT images using a modified Bayesian residual transform. *Biomedical Optics Express*, 9(5):2394–2406, 2018.
- J. Wang, M. Betke, and J. P. Ko. Pulmonary fissure segmentation on CT. *Medical Image Analysis*, 10(4):530–547, 2006.

# Dynamic Pacemaker Artifact Removal (DyPAR) from CT Data using CNNs

**Tanja Lossau (née Elss)<sup>1,2</sup>**

TANJA.LOSSAU@PHILIPS.COM

**Hannes Nickisch<sup>1</sup>**

HANNES.NICKISCH@PHILIPS.COM

**Tobias Wissel<sup>1</sup>**

TOBIAS.WISSEL@PHILIPS.COM

**Samer Hakmi<sup>3</sup>**

S.HAKMI@UKE.DE

**Clemens Spink<sup>4</sup>**

C.SPINK@UKE.DE

**Michael M. Morlock<sup>2</sup>**

MORLOCK@TUHH.DE

**Michael Grass<sup>1</sup>**

MICHAEL.GRASS@PHILIPS.COM

<sup>1</sup> Philips Research, Hamburg, Germany    <sup>2</sup> Hamburg University of Technology, Germany

<sup>3</sup> University Heart Center Hamburg, Germany    <sup>4</sup> Department of Diagnostic and Interventional Radiology and Nuclear Medicine, University Medical Center Hamburg-Eppendorf, Germany

## Abstract

Metal objects in the human heart like implanted pacemakers frequently occur in elderly patients. Due to cardiac motion, they are not static during the CT acquisition and lead to heavy artifacts in reconstructed CT image volumes. Furthermore, cardiac motion precludes the application of standard metal artifact reduction methods which assume that the object does not move. We propose a deep-learning-based approach for dynamic pacemaker artifact removal which deals with metal shadow segmentation directly in the projection domain. The data required for supervised learning is generated by introducing synthetic pacemaker leads into 14 clinical data sets without pacemakers. CNNs achieve a Dice coefficient of 0.913 on test data with synthetic metal leads. Application of the trained CNNs on eight data sets with real pacemakers and subsequent inpainting of the post-processed segmentation masks leads to significantly reduced metal artifacts in the reconstructed CT image volumes.

**Keywords:** Cardiac CT, Metal Artifact Reduction, Convolutional Neural Networks

## 1. Introduction

High-density objects like metallic devices lead to streak-shaped artifacts in reconstructed CT images which significantly degrade the image quality and the diagnostic value, see Figure 1. These artifacts are mainly caused by inconsistencies in the projection data due to effects like beam hardening and photon starvation (Mouton et al., 2013). Therefore, standard procedures for metal artifact reduction (MAR) are based on sinogram inpainting by treating metal-affected values as missing data. They comprise the following steps: 1) segmentation of the metal in an initially reconstructed image volume 2) forward projection of the metal mask to yield the metal shadow in the sinogram 3) inpainting of the metal shadow 4) reconstruction of the improved image volume with reinserted metal mask from step 1). A multitude of variants of this method is known (Kalender et al., 1987; Meyer et al., 2010), but all of them assume that the object is static during acquisition.

In case of implanted pacemakers, motion aggravates the metal artifacts and additionally hampers the evaluation of neighboring anatomy for example with regard to inflammation or calcification. As

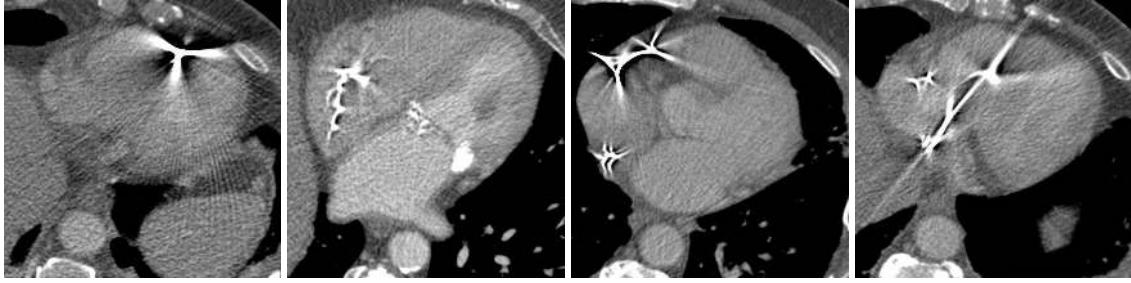


Figure 1: Pacemakers may lead to heavy streak-shaped artifacts in cardiac CT images. Especially in ungated CT data, motion aggravates artifacts caused by electrodes and leads.

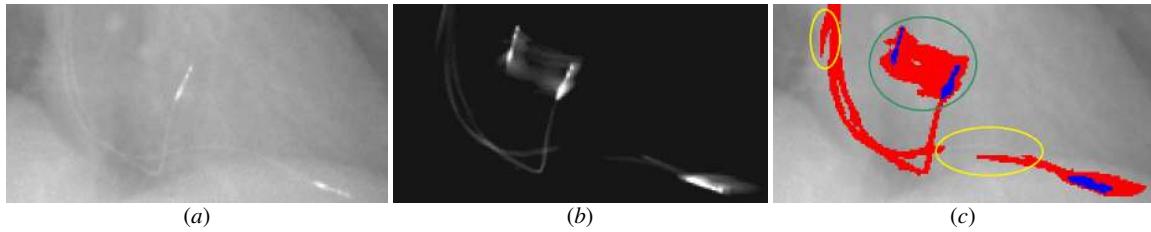


Figure 2: Standard MAR procedures are not applicable for moving metal objects like pacemakers. (a) Projection view which contains the metal shadow of two pacemaker leads and electrodes. (b) The corresponding forward projected pacemaker shadow is perturbed by cardiac motion artifacts. (c) Duplicated electrode shadows (green circle) as well as shifted and interrupted lead shadows (yellow circles) preclude conventional sinogram inpainting.

illustrated in Figure 2, standard MAR procedures fail in the presence of motion due to inconsistencies between the real metal shadow and the estimated one (see Figure 2b) which is obtained by step 2) of the standard MAR procedure.

In this paper, an alternative approach for dynamic MAR is investigated which deals with deep-learning-based segmentation of the metal shadow directly in projection data. Convolutional neural networks (CNNs) have been driving advances in several pattern recognition and semantic segmentation tasks (Krizhevsky et al., 2012; Ronneberger et al., 2015). Also in the field of CT MAR they are increasingly used, e.g. for sinogram correction (Gjesteby et al., 2017) and image-based artifact reduction (Zhang and Yu, 2018; Xu and Dang, 2018). In contrast to these methods, our dynamic pacemaker artifact removal (DyPAR) approach does not rely on initially reconstructed image volumes which are potentially motion perturbed, i.e. it is applicable to static and non-static objects.

We use a forward model to synthesize pacemaker lead projection data and train CNNs to segment the metal directly in the projection domain. Inpainting of the post-processed segmentation masks and subsequent reconstruction yield CT image volumes with reduced artifacts. The generalization capabilities and transferability to MAR in clinical practice are investigated based on eight test cases with real pacemakers and two additional test cases without pacemaker.

## 2. Data

### 2.1. Synthetic learning data

Section 3.1 details the generation of projection data with synthetic pacemaker leads and corresponding segmentation masks for subsequent network training. The following described *reference data with pacemakers* and *target data without pacemakers* form the basis of this synthetic learning data. Leads are inserted into the *target data without pacemakers*, whereby reasonable lead positions and pathways are extracted from the *reference data with pacemakers*.

**Reference data with pacemakers** Seven reconstructed CT image volumes with pacemakers are collected for the extraction of pacemaker lead positions with respect to the cardiac anatomy. Dual as well as triple chamber pacemakers are included, i.e. synthesis of right atrial leads, right ventricular leads and coronary sinus leads is aimed for.

**Target data without pacemakers** Synthetic leads are inserted into 14 contrast-enhanced clinical cardiac CT data sets without pacemakers. The reconstructed image volumes as well as the corresponding raw projection data are required for the insertion process. In all target cases, acquisition was performed with a 256-slice CT scanner (Brilliance iCT, Philips Healthcare, Cleveland, OH, USA) using a retrospective gating protocol without dose modulation. The helical trajectory exhibits a gantry rotation speed of 0.272 sec per turn. The pitch and the number of recorded projection views per turn vary between 0.16/0.18 and 1800/2400, respectively.

### 2.2. Clinical test data

In order to investigate generalization capabilities of the trained networks and the proposed DyPAR approach in clinical practice, ten additional cardiac CT data sets (i.e. the reconstructed image volume and the corresponding raw projection data) are collected. Eight data sets exhibit real pacemakers and allow one to evaluate the robustness of the CNNs with regard to variations in contrast-enhancement, motion levels, acquisition settings (pitch, w/wo dose modulation, w/wo gating) and lead pathways. Furthermore, the networks behavior in the presence of unseen features like electrodes or defibrillators is investigated. The remaining two test data sets without pacemakers are considered to quantify false positives. Both exhibit severe calcifications at the coronary arteries and the aortic valve. Furthermore, sternal steel wires for median sternotomy closure and a stent at the left main artery are present in one of these *no-pacemaker data sets*. The evaluation results are presented in Section 4.1.

## 3. Method

CNNs are trained for the task of segmenting pacemaker metal shadows in the projection domain. The required data for supervised learning is generated using a forward model for synthetic lead introduction. The Subsections 3.1 and 3.2 detail the data generation and the supervised learning processes. Finally, the trained networks are integrated into the DyPAR pipeline for application on clinical test data as described in Subsection 3.3.

### 3.1. Learning data generation

The data generation process is visualized in Figure 3 and requires one reference case with pacemaker and one target case without pacemaker. For both CT image volumes, corresponding heart meshes

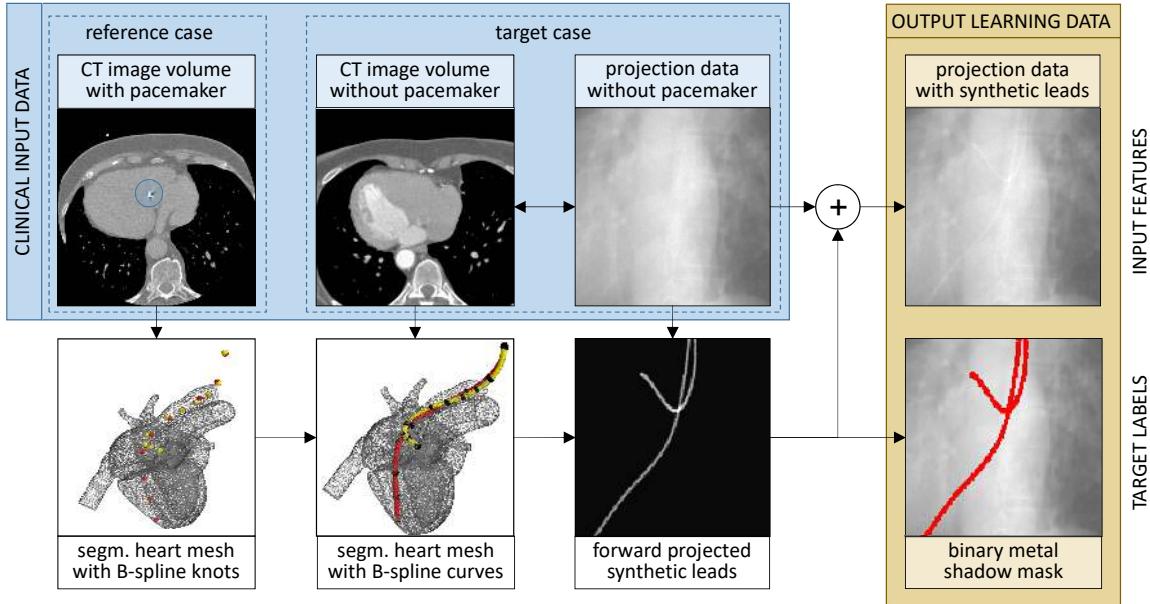


Figure 3: Synthetic leads are introduced into the projection data of clinical cases without pacemakers. Consideration of the segmented heart meshes allows sensible insertion.

are obtained by model-based heart segmentation according to (Ecabert et al., 2008). At least ten B-spline knots are manually selected along each pacemaker lead in the reference case. Based on the point-to-point correspondences in the segmented heart meshes, thin plate spline smoothing is performed to transform the B-spline knots from the reference case into the target case. The corresponding B-spline curve is determined by cubic B-spline interpolation for each pacemaker lead, separately. Dilation of the resulting lines with a chosen lead diameter of 2 millimeters and an attenuation value of 4500 HU yields a binary lead mask in the target image geometry. Subsequent forward projection delivers the corresponding metal shadow in the originally acquired projection geometry of the target case. The required input data for supervised learning is finally obtained by summation of the original projection data and the forward projected lead mask. Thresholding with zero defines the corresponding target segmentation mask.

### 3.2. Supervised learning

The data generation process is applied two times per reference case as twice as many target cases are available. An average number of 10 000 views per case is included in the supervised learning process. Each view contains  $128 \times 672$  line integrals determined by the detector height and width of the Brilliance iCT. On average, 1.4% of the projection data contains object voxels. The data is case-wise separated into the subsets training, validation and testing with a ratio of 6:4:4, or rather 3:2:2 with respect to the corresponding reference cases. By this data separation strategy, it is ensured that pacemaker geometries and background line integrals are disjoint among the subsets.

**Patch sampling** During training, the CNN takes 3D patches of size  $100 \times 100 \times 11$  voxels as input and delivers 2D patches of size  $20 \times 20$  as output. The first and second dimension of the network input contain the information of the detector row and column, while the third dimension

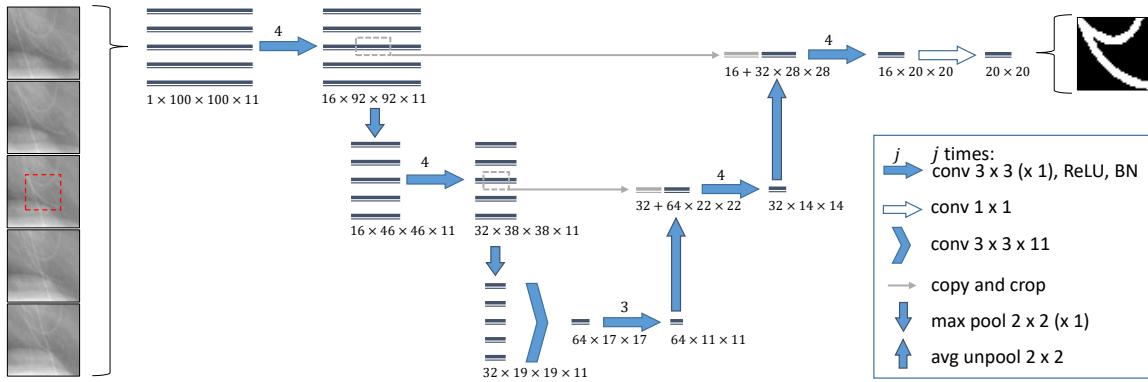


Figure 4: Adapted U-net architecture for 2D lead segmentation from multi-view input patches. Each box corresponds to a multi-channel feature map. For clarity only five instead of eleven input views are illustrated. The actual shape of each feature map is denoted at the lower left edge of the box. The arrows represent the different operations.

corresponds to the projection view. The third patch dimension is equidistantly sampled with respect to the number of views per gantry turn in such a way that 12 degrees gantry rotation are captured. Inclusion of neighboring views has the benefit, that the network gets additional information about the rotation velocity of supposed pacemaker leads, i.e. information about the distance to the rotation center. The target patch corresponds to the center of the sixth view. In order to compensate the extreme foreground-background class imbalance, the sampling process is controlled in such a way that 75% of the training patches contain at least one object voxel. The remaining 25% are randomly sampled.

**Network architecture** Figure 4 shows the utilized U-Net architecture adapted from (Ronneberger et al., 2015). In the contracting path, 2D lead features are extracted for each view, separately. These features are joint in the bottleneck to exploit the temporal information. In the expanding path, merely location information of the center slice to be segmented are copied and cropped from the contracting path. The slim architecture with its shared weights in the contracting path has a relatively low number of 423 730 learned parameters. The networks output is reduced in the first two dimensions, as no internal padding is performed. The fully convolutional network design allows for arbitrary output shapes. During validation and testing, the full detector size of  $128 \times 672$  voxels is segmented in a single step by previously enlarging the input projection data sets with respect to the networks receptive field of  $11 \times 81 \times 81$  pixels using symmetric padding. During parameter tuning we observed, that in particular the receptive field size and the number of projection views per patch are crucial to the network performance.

**Learning setup** The stochastic gradient descent solver Adam (Kingma and Ba, 2015) with an initial learning rate of 0.01, a mini-batch size of 32 and a momentum of 0.8 is used for network optimization. Training is performed over 30 epochs while one epoch is defined by 10e5 processed samples. The learning rate decreases with a factor of two after every 10<sup>th</sup> epoch and L2 regularization with a weight of 0.0002 is used. The learning process is driven by the focal loss (Lin et al., 2018) with focusing parameter  $\gamma = 2$ .

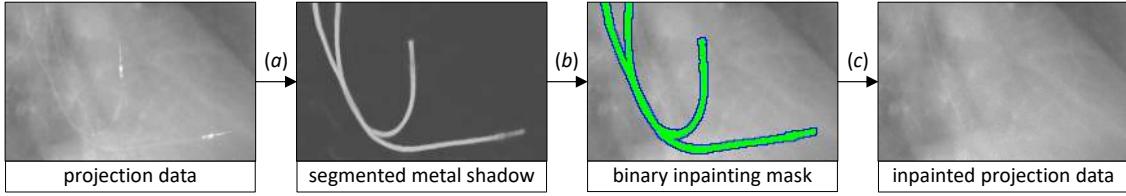


Figure 5: Schematic steps for sinogram correction. (a) The pacemaker metal shadow is segmented by CNNs. (b) The resulting probability map is transformed into a binary mask. (c) Metal-affected values are inpainted to yield pacemaker-free projection data.

**Bagging approach** An ensemble of five CNNs is learned using the aforementioned network architecture and hyper-parameter settings by the following bagging approach:

1. Four validation cases and four test cases are randomly sampled.
2. Network training is performed based on the remaining six clinical cases.
3. After every epoch, the networks generalization capabilities are examined by calculating the Dice coefficient with a threshold of 0.5 on the validation set.
4. The model with the highest validation metric within 30 epochs of training is selected for performance evaluation and application in the DyPAR pipeline.
5. Steps 1.-4. are performed five times in total.

### 3.3. DyPAR pipeline

The proposed dynamic pacemaker artifact removal (DyPAR) approach deals with sinogram correction and subsequent filtered backprojection ([Koken and Grass, 2006](#)). As illustrated in Figure 5, the sinogram correction comprises three steps:

**(a) Metal shadow segmentation** In order to increase the robustness of the metal shadow segmentation, the entire ensemble of five CNNs yielded by the bagging approach is applied on the previously padded input raw projection data. It has to be noted that merely projection data which is required for cardiac field-of-view reconstruction is processed. That means, the pulse generator below the clavicle, for instance, is not segmented. The output probability map is averaged across the ensemble and contains values between 0 and 1. In contrast to standard MAR approaches, the segmentation delivers probability maps which are independent of the 3D motion blur and exactly located on the real metal shadows.

**(b) Post-processing** By thresholding the probability map with 0.15 and extracting the largest connected component using a  $3 \times 3 \times 3$  structure element, a binary inpainting mask is determined. The choice of the relatively low threshold 0.15 is motivated by the fact that incomplete metal masks may introduce new artifacts after inpainting and reconstruction. For the task of MAR, false positives are less severe than false negatives, i.e. the sensitivity is a more important performance measure than the specificity.

Table 1: Mean network performance measures (in percent) on training, validation and testing subsets are compared for different post-processing settings, i.e. thresholds  $t$  for separation of object and background voxels and optional largest connected component (LCC) extraction.

subset $t / \text{LCC}$	Dice coefficient			sensitivity			specificity			AUC X
	0.5 X	0.15 X	0.15 ✓	0.5 X	0.15 X	0.15 ✓	0.15 X	0.5 X	0.15 ✓	
testing	91.27	79.44	84.92	89.63	98.05	97.96	99.90	99.28	99.50	99.86
validation	90.93	78.40	84.23	90.04	98.25	97.88	99.90	99.31	99.54	99.86
training	94.55	80.22	86.32	95.48	99.89	99.87	99.92	99.36	99.98	99.98

**(c) Inpainting** The predicted metal-affected values (green voxels in Figure 5) are treated as missing data. The projection data is filled by distance weighted 2D linear interpolation, based on surrounding line integrals (blue voxels in Figure 5). Of course, alternative inpainting approaches instead of view-wise linear interpolation are possible. The focus here is however on the metal shadow segmentation step. Filtered backprojection (Koken and Grass, 2006) of the inpainted projection data finally delivers the CT image volume without pacemaker and concomitant artifacts.

## 4. Experiments and results

For all experiments, the Microsoft Cognitive Toolkit (CNTK v2.5+, Microsoft Research, Redmond, WA, USA) is used as deep learning framework. Table 1 summarizes networks performance measures during the five-fold bagging approach. Dice coefficients, sensitivity, specificity and the area under curve (AUC) are considered. These quantitative results illustrate that the CNNs are able to identify synthetic pacemaker leads in clinical projection data. To further evaluate generalization capabilities of the bagging ensemble in clinical practice, DyPAR is applied on the *clinical test data* described in Section 2.2.

### 4.1. Evaluation on clinical test data

In order to identify strengths and weaknesses of the trained neural networks, qualitative evaluations of the segmentation masks and the output image volumes are performed by visual inspection. For comparison, image-based metal shadow extraction is considered, whereby the metal masks in the image domain are segmented using 3D hysteresis thresholds of 1000 HU and 1500 HU. The metal shadow areas are yielded by forward projection and thresholding with zero.

Both metal shadow extraction approaches, the image-based one and ours, are compared in Figure 6. For both methods, few false negatives are observable (see Figure 6c,h,j,k). In contrast to the image-based approach, our networks deliver segmentation results which are not affected by blurring between different motion states (see Figure 6a,g,h,j). Despite the lag of dedicated learning data, the networks object-background separation also holds for electrodes (see Figure 6a-k) and defibrillators (see Figure 6b,f). Misinterpretation of spine as lead is identified as one possible error source (see Figure 6f), whereas metal shadow kinks (see Figure 6d) and moderate noise do not seem to confuse the neural networks. More severe noise caused by dose modulation is indeed a limiting factor, but projection views recorded with reduced radiation doses are generally not taken into account during back projection. In cases with dose modulation, the post-processing step has to be adapted in such a way that the largest connected component is determined for each high dose segment, separately.

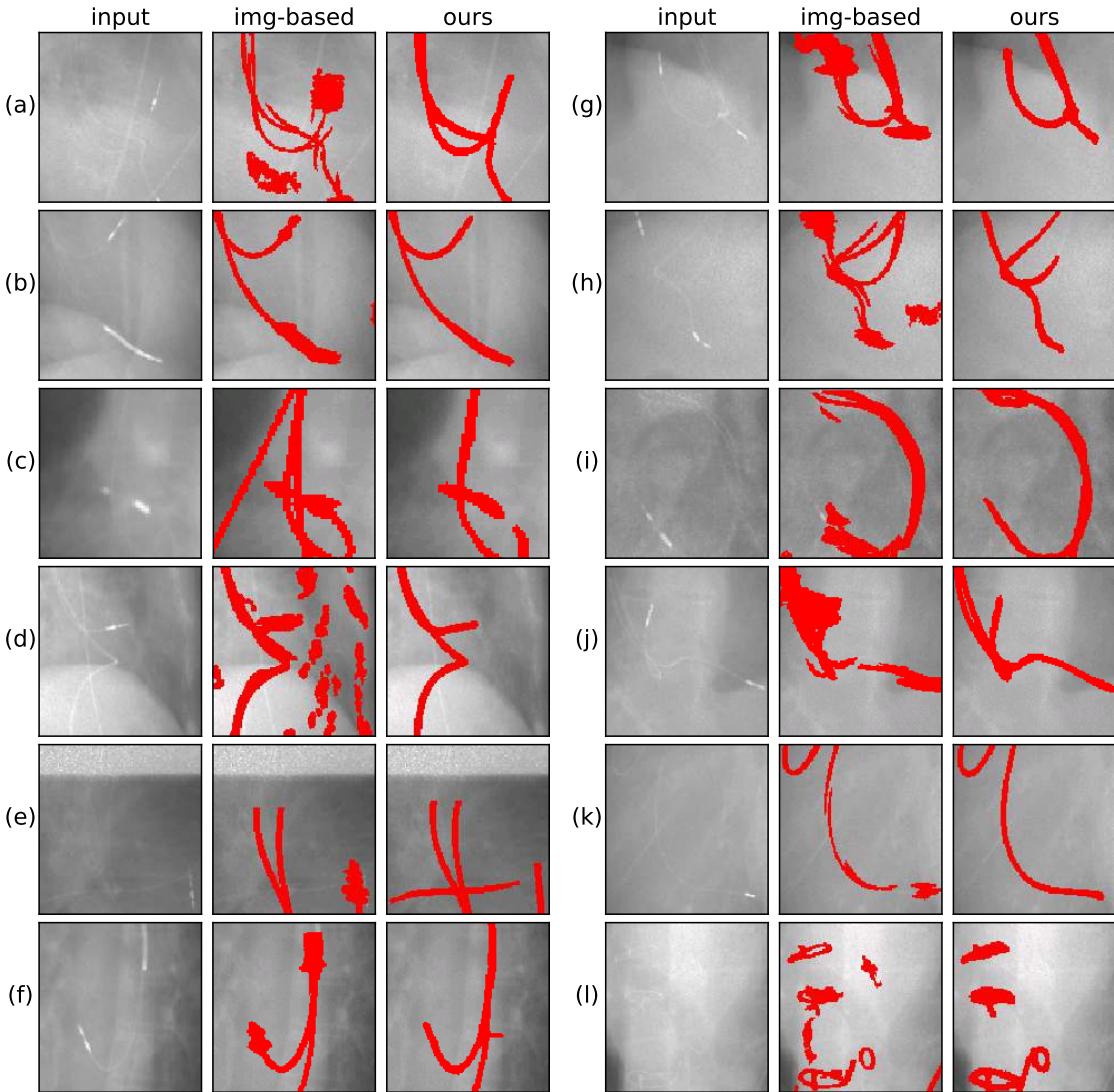


Figure 6: Example projections with real metal implants and corresponding metal shadow areas predicted by the image-based approach and during our DyPAR pipeline.

Based on a single projection view, ECG and pacemaker leads are visually hard to distinguish. The CNNs are remarkably successful in their differentiation and seem also to consider the rotation velocities (see Figure 6a,c,d). This strength is restricted in case of horizontally aligned ECG leads (see Figure 6e) as for these segments, the rotation velocity can not be measured locally. False positives are increasingly present in the image-based metal shadows, e.g. caused by severe calcifications, ECG-leads and bone (see Figure 6a-c,d,h). In the two *no-pacemaker data sets*, false positive rates of 0.1125% and 3.3297% are achieved by the networks. False positive activations are caused by parts of the ECG leads and the sternal steel wires (see Figure 6l). In contrast to the image-based approach which exhibits false positive rates of 0.8647% and 3.6866%, calcifications and stents are not

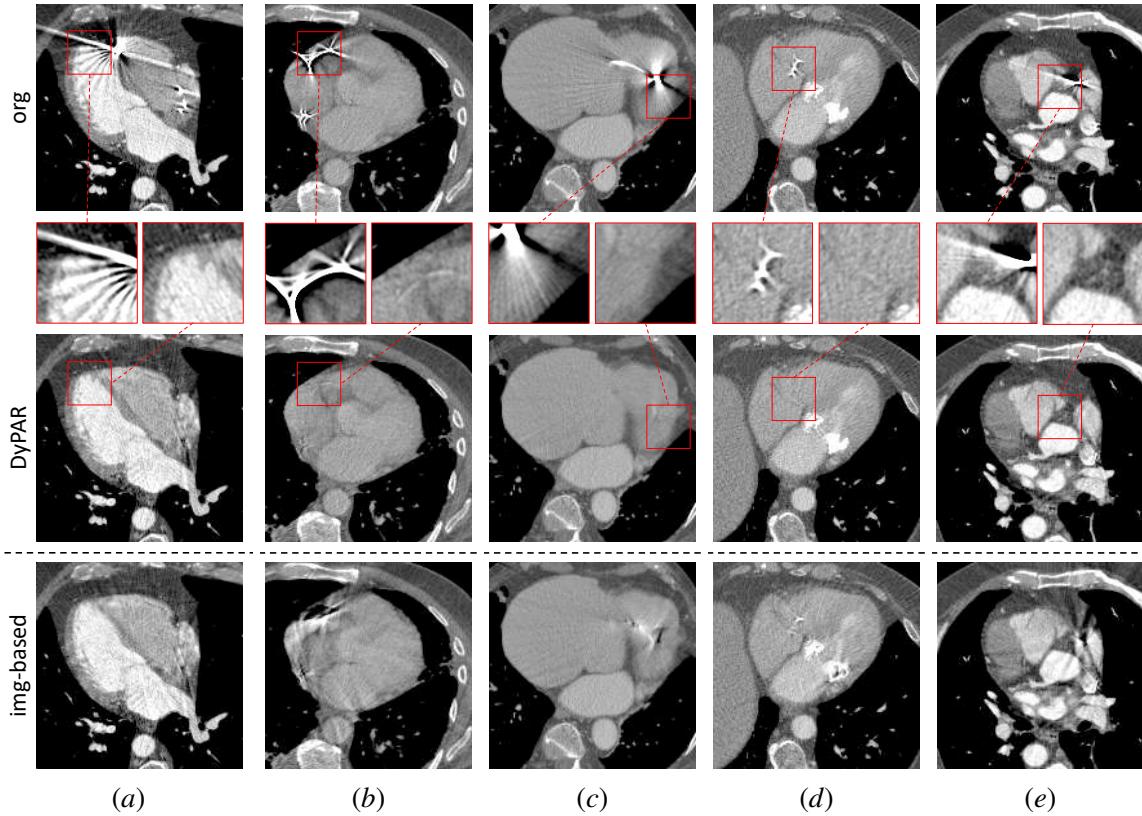


Figure 7: Axial slices of clinical test cases before and after DyPAR. For comparison, the corresponding results of the image-based MAR approach are considered.

misinterpreted by the networks. The superiority of our method over the image-based is observable in the projection and the image domain.

Figure 7 shows some results of the DyPAR with significant reduction of the streak-shaped metal artifacts. Especially in the zoomed regions, evaluation of the neighboring anatomy is facilitated. The image-based approach is in contrast not robust regarding cardiac motion and leads to increased blurring in the neighboring anatomy (see Figure 7a), incomplete metal removal (see Figure 7c,d) and introduction of new severe artifacts (see Figure 7b,e). In fact, our method also introduces new artifacts whenever the inpainting causes inconsistencies in the projection data (see Figure 7a, right). Nevertheless, the DyPAR approach demonstrates great generalization capabilities as it is transferable to different scanner types (also tested: Brilliance 64, Philips Healthcare, Cleveland, OH, USA) and ungated CT data sets (see Figure 7b-d). In these cases, contrast enhancement and acquisition settings like the pitch and the gantry rotation speed vary from the learning data.

## 5. Discussion and Conclusion

Establishment or maintenance of consistent projection data is essential for artifact removal or avoidance. As already mentioned in Subsection 3.3, simple 2D linear interpolation should not be the method of choice as edge preservation is not ensured. Since neural networks also showed great results in image inpainting (Yang et al., 2017), a deep-learning-based approach should be considered. Furthermore, metal reinsertion with respect to the cardiac motion is still an issue. So far, the metal is completely removed from the reconstructed CT image volumes.

The proposed data generation process is in principle extendable to different scanner types, acquisition modes and high-density objects. It enables on-site training for protocol-specific MAR. Hence, generalization to step-and-shoot data and transferability to other metal implants like artificial valves or metal clips will be part of future research.

To conclude, dynamic metal artifact removal based on metal shadow segmentation in the projection domain is feasible. CNNs trained on clinical data with synthetically introduced pacemaker leads show great generalization capabilities in the segmentation of electrode and defibrillator shadows. Quantitative validation studies are required to assess the transferability of these promising initial results to pacemaker CT artifact reduction in clinical practice.

## References

- Olivier Ecabert, Jochen Peters, Hauke Schramm, Cristian Lorenz, Jens von Berg, Matthew J Walker, Mani Vembar, Mark E Olszewski, Krishna Subramanyan, Guy Lavi, et al. Automatic model-based segmentation of the heart in CT images. *IEEE Transactions on Medical Imaging*, 27(9):1189–1201, 2008.
- Lars Gjesteby, Qingsong Yang, Yan Xi, Ye Zhou, Junping Zhang, and Ge Wang. Deep learning methods to guide CT image reconstruction and reduce metal artifacts. In *Medical Imaging 2017: Physics of Medical Imaging*, volume 10132, page 101322W. International Society for Optics and Photonics, 2017.
- Willi A Kalender, Robert Hebel, and Johannes Ebersberger. Reduction of CT artifacts caused by metallic implants. *Radiology*, 164(2):576–577, 1987.
- Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- P Koken and M Grass. Aperture weighted cardiac reconstruction for cone-beam CT. *Physics in Medicine and Biology*, 51(14):3433, 2006.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.
- Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- Esther Meyer, Rainer Raupach, Michael Lell, Bernhard Schmidt, and Marc Kachelrieß. Normalized metal artifact reduction (NMAR) in computed tomography. *Medical Physics*, 37(10):5482–5493, 2010.

Andre Mouton, Najla Megherbi, Katrien Van Slambrouck, Johan Nuyts, and Toby P Breckon. An experimental survey of metal artefact reduction in computed tomography. *Journal of X-ray Science and Technology*, 21(2):193–226, 2013.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–231. Springer, 2015.

Shiyu Xu and Hao Dang. Deep residual learning enabled metal artifact reduction in CT. In *Medical Imaging 2018: Physics of Medical Imaging*, volume 10573, page 105733O. International Society for Optics and Photonics, 2018.

Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3, 2017.

Yanbo Zhang and Hengyong Yu. Convolutional neural network based metal artifact reduction in x-ray computed tomography. *IEEE transactions on medical imaging*, 37(6):1370–1381, 2018.

# Group-Attention Single-Shot Detector (GA-SSD): Finding Pulmonary Nodules in Large-Scale CT Images

**Jiechao Ma**<sup>1,3</sup>

MAJCH7@MAIL2.SYSU.EDU.CN

**Xiang Li**<sup>1</sup>

LIXIANG651@GMAIL.COM

**Hongwei Li**<sup>2</sup>

HONGWEI.LI@TUM.DE

**Bjoern H Menze**<sup>2</sup>

BJOERN.MENZE@TUM.DE

**Sen Liang**<sup>3</sup>

LSEN@INFERVISION.COM

**Rongguo Zhang**<sup>3</sup>

ZRONGGUO@INFERVISION.COM

**Wei-Shi Zheng**<sup>1</sup>

WSZHENG@IEEE.ORG

<sup>1</sup> School of Data and Computer Science, Sun Yat-sen University, China.

<sup>2</sup> Department of Computer Science, Technical University of Munich, Germany.

<sup>3</sup> Infervision Inc, China.

## Abstract

Early diagnosis of pulmonary nodules (PNs) can improve the survival rate of patients and yet is a challenging task for radiologists due to the image noise and artifacts in computed tomography (CT) images. In this paper, we propose a novel and effective abnormality detector implementing the attention mechanism and group convolution on 3D single-shot detector (SSD) called group-attention SSD (GA-SSD). We find that group convolution is effective in extracting rich context information between continuous slices, and attention network can learn the target features automatically. We collected a large-scale dataset that contained 4146 CT scans with annotations of varying types and sizes of PNs (even PNs smaller than 3mm). To the best of our knowledge, this dataset is the largest cohort with relatively complete annotations for PNs detection. Extensive experimental results show that the proposed group-attention SSD outperforms the conventional SSD framework as well as the state-of-the-art 3DCNN, especially on some challenging lesion types.

**Keywords:** Lung Nodule Detection, Single Shot Detector, Attention Network, Group Convolution

## 1. Introduction

Lung cancer continues to have the highest incidence and mortality rate worldwide among all forms of cancers (Bray et al., 2018). Because of its aggressive and heterogeneous nature, diagnosis and intervention at the early stage, where cancer manifests as pulmonary nodules, are vital to survival (Siegel and Jemal, 2018). Although the use of a new generation of CT scanners improves the detection of pulmonary nodules, certain nodules (such as ground-glass nodules, GGN) are still misdiagnosed due to noise and artifacts in CT imaging. (Manning et al., 2004; Hossain et al., 2018). The design of a reliable detection system is increasingly needed in clinical practice.

Deep learning techniques using convolution neural networks (CNN) is a promising and effective approach to assisting lung nodule management. For example, Setio (Setio et al., 2016) proposed a system for pulmonary nodule detection based on multi-view CNN, where the network is fed with nodule candidates rather than whole CT scans. Wang (Wang et al., 2018a) presented a 3D CNN

model trained with feature pyramid networks (FPN) (Lin et al., 2016) and achieved the state-of-the-art on LUNA16<sup>1</sup>. However, all these algorithms neither make use of the spatial attention across the neighboring slices nor introduce the attention mechanism for region of interest, because the regional distribution of target PNs and the non-PNs is highly unbalanced. Therefore, learning to automatically weight the importance of slices and pixels is essential in pulmonary nodules detection.

In this work, to address the problem of indiscriminate weighting of pixels and slices, we propose a lung nodule detection model called group-attention SSD (GA-SSD), which leverages one-stage single-shot detector (SSD) framework (Liu et al., 2016; Fu et al., 2017a; Luo et al., 2017) and attention module with group convolutions. Firstly, a group convolution is added at the beginning of the GA module to weight the importance of input slices. Secondly, the attention mechanism is integrated into the grouped features to enhance the weight of nodule’s pixels on a 2D image.

We evaluate the proposed system on our challenging large-scale dataset containing 4,146 patients. Different from existing datasets, the cohort contains eight categories of PNs including ground-glass nodules (GGNs) which are hard-to-detect lesions of clinical significance yet not usually included in conventional datasets.

## 2. Related Works

**Object Detection.** Recent object detection models can be grouped into one of two types (Liu et al., 2018), two-stage approaches (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015) and one-stage methods (Redmon et al., 2016; Liu et al., 2016). The former generates a series of candidate boxes as proposals by the algorithm and then classifies the proposals by convolution neural network. The latter directly transforms the problem of target border location into a regression problem without generating candidate boxes. It is precise because of the difference between the two methods, the former is superior in detection accuracy and location accuracy, and the latter is superior in algorithm speed.

**Attention Modules.** The inspiration of attention mechanism comes from the mechanism of human visual attention. Human vision is guided by attention which gives higher weights on objects than background. Recently, attention mechanism has been successfully applied in natural language processing (Vaswani et al., 2017; Cho et al., 2014; Sutskever et al., 2014; Yang et al., 2016; Yin et al., 2015) as well as computer vision (Fu et al., 2017b; Zheng et al., 2017; Sun et al., 2018). Most of the conventional methods which solve the object detection problems neglect the correlation between proposed regions. The Non-local Network (Wang et al., 2018b) and the Relation networks (Hu et al., 2018) were translational variants of the attention mechanism and utilize the interrelationships between objects. In medical image analysis community, oktay (Oktay et al., 2018) introduced attention mechanism to solve the pancreas segmentation problem. Our method is motivated by these works, aiming at medical images, to find the inter-correlation between CT slices and between lung nodule pixels.

**Group Convolution.** Group convolution first appeared in AlexNet(Krizhevsky et al., 2012). To solve the problem of insufficient memory, AlexNet proposed that the group convolution approach could increase the diagonal correlation between filters and reduce the training parameters. Recently, many successful applications have proved the effectiveness of group convolution modules such as channel-wise convolution including the Xception (Szegedy et al., 2015, 2016) (Extreme Inception) and the ResNeXt(Xie et al., 2017).

---

1. <https://luna16.grand-challenge.org/>

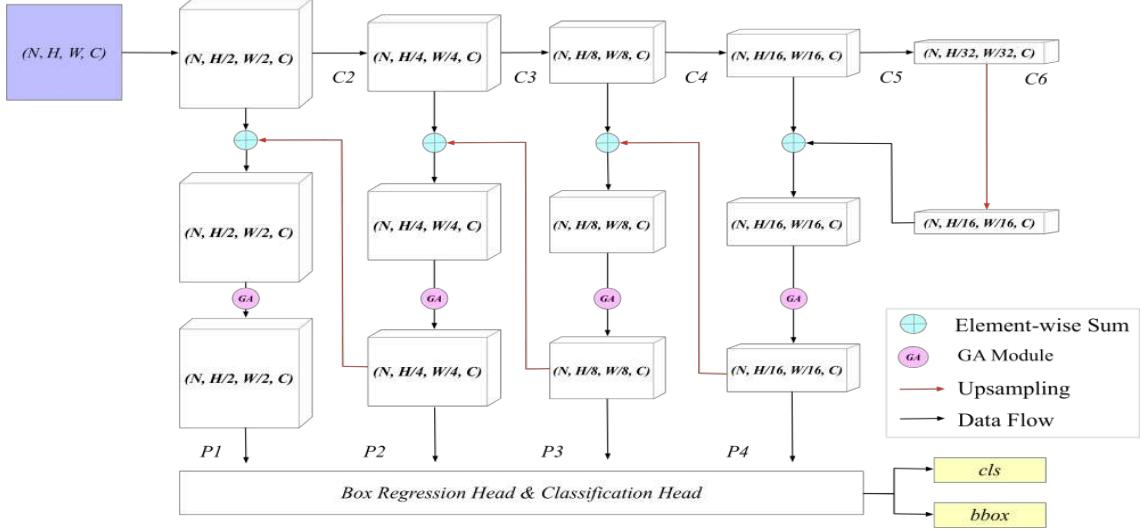


Figure 1: The architecture of SSD framework with FPN and attach the GA module.

### 3. Methodology

In this section, we present an effective 3D SSD framework for lung nodule detection because the detection of lung nodules relies on 3D information. The proposed framework has two highlights: the attention module and grouped convolution. We call this new model group attention SSD (GA-SSD) because we integrate group convolution with attention modules.

#### 3.1. Overall Framework

The proposed 3D SSD shares the basic architecture of the classic SSD (Liu et al., 2016). And the network structure of the GA-SSD can be divided into two parts: the medical image loading sub-network for reading CT scans (pre-process) and the backbone sub-network for feature extraction. Specifically, we use the deeply supervised ResNeXt (He et al., 2016) structure as the backbone, and add GA structure to both pre-process sub-network and backbone sub-network respectively.

#### 3.2. SSD Architectures

The classic SSD network was established with VGG16 (Simonyan and Zisserman, 2014). Compared with Faster RCNN (Ren et al., 2015), the SSD algorithm does not generate proposals, which greatly improves the detection speed (to handle the large-scale dataset). The basic idea of SSD is to transform the image into different sizes (image pyramid), detect them separately, and finally synthesize the results.

In this work, the original ResNeXt structure is modified into FPN-like layers (Figure 1). In order to detect the small object, four convolution layers (P1, P2, P3, P4) are added to construct the different size candidate boxes. And these outputs (boxes) are optimized by two different losses (regression, classification), resulting in one class confidence (each default box generates several class confidences) and one output regression localization (each default box generates four coordinate values  $(x, y, w, h)$ ). In order to utilize the three-dimensional information between lung CT slices, we

modify the basic structure of SSD network to improve the performance of the detection framework. The backbone network uses 3D convolution with the same padding instead of the conventional 2D convolution layer. And the Rectified linear unit (ReLU) is employed as activation functions to the nodes. Additionally, we apply dropout regularization([Srivastava et al., 2014](#)) to prevent complex node connections.

### 3.3. GA Modules

To imitate the usual viewing habits of radiologists who usually screen for nodule lesions in 2-D axial plane at the thin-section slice, we propose a new medical imaging group convolution and attention-based network (GA module (Figure 2)) to tell the model which slices of the patient to focus on and automatically learn the weight of these slices. See figure 2, assume that the input feature map is  $(N, H, W, C)$ , which means the channel is  $C$ , the batch size is  $N$ , the width and height are  $H, W$ , respectively. Suppose the number of group convolutions is  $M$  (we use  $M = 9$  by default). So the operation of this group convolution is to divide channels into  $M$  parts. Each group corresponds to  $C/M$  channels and is convoluted independently. After each group convolution is completed, the output of each group is concatenated as the output channel  $(N, H, W, C)$ . And the attention mechanism based on sequence generation can be applied to help convolutional neural networks to focus on non-local information of images to generate weighted inputs of the same sizes as the original inputs.

The GA behavior in Eq.(1) is due to the fact that all pixels are considered in the operation.  $f(x_i, x_j)$  is used to calculate the pairwise relationship between target  $i$  and all other associated pixel  $j$ . This relationship is as follows: the farther the pixel distance between  $i$  and  $j$  is, the smaller the  $f$  value is, indicating that the  $j$  pixel has less impact on  $i$ .  $g(x)$  is used to calculate the eigenvalues of the input signal at the  $j$  pixel.  $C(x)$  is a normalized parameter. In figure 2, we use three  $1 \times 1$  convolution layer to get corresponding features. Then use the softmax function ( $f(x)$ ) and gaussian function ( $g(x)$ ) to get the attention information.

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j). \quad (1)$$

Our approach improves parameter efficiency and adopts group convolution and attention module, where group convolution acts to find the best feature maps (i.e., highlight several slices from the input CT scans), and the attention module acts to find the location of the nodule (i.e., the size and shape of the target nodule in a specific feature map).

In addition, due to the simplicity and applicability of GA module, it can be easily integrated into a standard CNN architecture. For example, we apply this module not only to the data loading sub-network but also the feature extraction stage of the network, which allows the model to automatically and implicitly learn some correlated regions of different features, and focuses on areas that the model needs to focus on.

## 4. Experiments and Results

### 4.1. A Large-scale Computed Tomography Dataset

A cohort of 4146 chest helical CT scans was collected from various scanners from several centers in China. Each chest CT scan contains a sequence of slices. Pulmonary nodules were la-

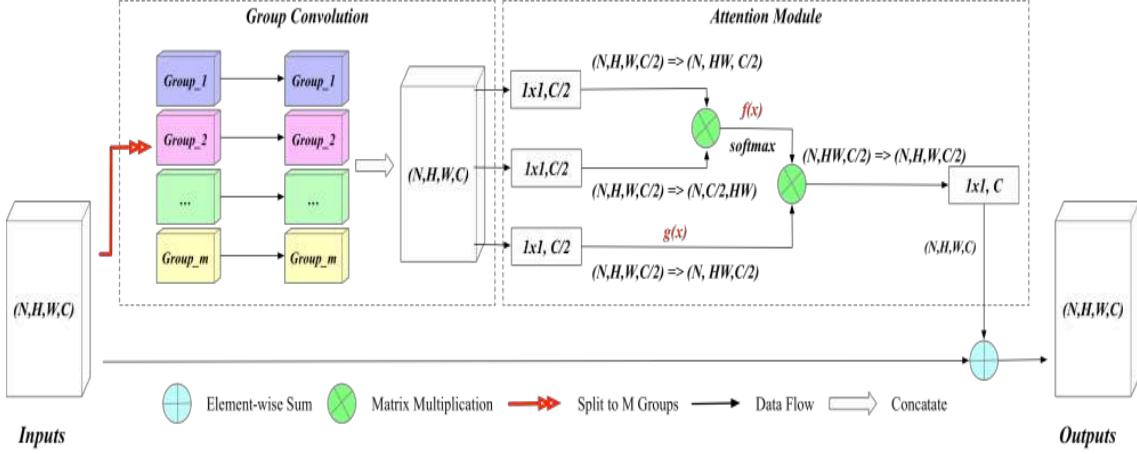


Figure 2: The architecture of GA module, it can be divided into two parts: The former enforces a sparsity connection by partitioning the inputs (and outputs) into disjoint groups. The latter use the concatenated groups to find the non-local information.

beled by experienced radiologists after evaluating their appearance and sizes. They are divided into eight categories: *calcified nodule with two different sizes*, *pleural nodule with two different sizes*, *solid nodules two different sizes*, *ground-glass nodule divided into pure-GGN and sub-solid nodules (mixed-GGN)* as shown in (Figure 3). To the best of our knowledge, the current dataset is the largest cohort for PNs detection and with eight categories and varied sizes of annotated PNs. The detection of ground-glass nodules is important and challenging in clinical practice; however, it is not included as a part of the PNs detection task in conventional datasets such as LUNA16.

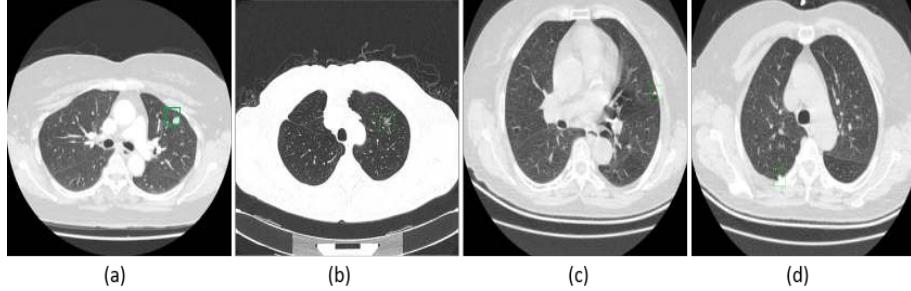


Figure 3: Sample CT images of the dataset used in evaluation of our deep learning model.(a) solid nodules, (b) subsolid nodules, (c) calcified nodules, and (d) pleural nodules

The images in our study were acquired by different CT scanners with Philips, GE, Siemens, Toshiba. All the chest CT images were acquired with the axial plane at the thin-section slice spacing (range from 0.8 to 2.5 mm). The regular dose CT scans were obtained at 100 kVp-140 kVp, tube current greater than 60mAs, 512 \* 512 pixel resolution; and the low-dose CT images were obtained that tube current less than 60mAs with all other acquisition parameters the same as those used to

obtain the regular dose CT. For our experiment, we randomly selected 80% of patients as training set and 20% of patients as testing set. Gradient updates were computed using batch sizes of 28 samples per GPU. All models were trained using the SGD optimizer, implemented batch normalization, and data augmentation techniques. Lung nodule detection performance was measured by CPM<sup>2</sup>. The evaluation is performed by measuring the detection sensitivity of the algorithm and the corresponding false positive rate per scan. This performance metric was introduced in (Setio et al., 2017).

#### 4.2. Effectiveness of GA Module for Data Load

Table 1 investigates the effects of different input methods. The baseline (multi-channel) used continuous slices as input and our approach added the GA module to this input mode. Using the GA module, our approach improved the CPM from 0.615 to 0.623. When we changed the 2.5D input (multi-channel 2D) to 3D volume input by reshaping the dimensions, the results improved by (+0.018) vs (+0.008). These results verified the effectiveness of the GA module for data pre-processing. In addition, our approach works better on 3D data than on 2D data.

Table 1: Comparision of the input method with GA module.

Input Method	2.5-D	3-D
Multi-channel	0.615	0.654
GA module	<b>0.623</b>	<b>0.672</b>

#### 4.3. Effectiveness of GA Module for Capturing Multi-scale Information

For detecting small objects, feature pyramid is a basic and effective component in the system. In general, small lung nodules are challenging for detectors, this is because there are numerous small nodules which are around  $20 \times 20$  pixels in the  $512 \times 512$  image data, making it difficult to localize. Thus, FPN is an important component in our framework for detecting small nodules.

In order to better investigate the impact of FPN on the detector, we conducted comparison experiments on a fixed set of feature layers. From Figure 1, we can see that on the basis of the original SSD (C2, C3, C4, C5, C6), the feature map of the latter layer uses upsampling to enlarge the size, and then adds with the former layer (original FPN). In the GA-FPN version, we use the GA module of the feature map as shown in Figure 2 to get the weight between feature maps. We choose to calculate the candidate box (like SSD) for several of the layers on the FPN. The lower layer, such as the P1 layer, has better texture information, so it is associated with good performance on small targets that can be identified by the detector. The higher level has stronger semantic information, and is associated with better results for the category classification of nodules. For the sake of simplicity, we do not share feature levels between layers unless specified.

Table 2 compares the effect of FPN and GA-FPN across the feature maps. According to Table 2 (left) results, using a three-layer (P4, P3, P2) model as a baseline, when a feature is used at a very early stage (like P1), it brought more false positives and harmed the performance (drops from 0.654

2. The code is opensource: <https://www.dropbox.com/s/wue67fg9bk5xdxt/evaluationScript.zip>

to 0.473). However, compared to the output of the P4 layer alone (0.680), the model with relatively lower information of the P3 layer gained better performance.

According to Table 2 (right) results, the framework improved the overall performance across feature layers. That the framework performance improved from 0.696 (best in FPN ) to 0.721 (best in GA-FPN) validated our conjecture that using the GA module could help the model learn more important feature layers.

Table 2: Comparision of the feature extraction with GA module.

Feature	FPN				GA-FPN			
P4	✓	✓	✓	✓	✓	✓	✓	✓
P3		✓	✓	✓		✓	✓	✓
P2			✓	✓			✓	✓
P1				✓				✓
CMP	0.680	<b>0.696</b>	0.654	0.473	<b>0.721</b>	0.703	0.672	0.554

#### 4.4. Comparison with State-of-the-art.

Extensive experiments were performed on our large CT dataset. We mainly compared our approach with current state-of-the-art methods for object detection in computer vision fields such as RCNN ([Ren et al., 2015](#); [He et al., 2016](#); [Xie et al., 2017](#)), YOLO ([Redmon et al., 2016](#)) and SSD ([Liu et al., 2016](#)) as well as current state-of-the-art method for PNs detection. The results are mainly summarized in Tabele 3 and the other detail components can be found as follow. From Table 3 <sup>3</sup>, we can observe that our system has achieved the highest CPM (0.733) with the fewest false positives rate (0.89) among this systems, which verifies the superiority of the improved GA-SSD in the task of lung nodule detection. On the classes of *p.ggn* and *m.ggn*, which are challenging to detect in clinical practice, our GA-SSD outperforms other approaches by a large margin.

To better justify the effectiveness of the proposed method, we conduct experiments over the LIDC-IDRI dataset ([Armato III et al., 2011](#)) and obtained the competitive result with the state-of-the-art ([Wang18 \(Wang et al., 2018a\)](#)) method (CPM scores: 0.863 vs 0.878).

## 5. Conclusion

In this paper, we proposed a novel group-attention module based on 3D SSD with ResNeXt as the backbone for pulmonary nodule detection with CT scans. The proposed model showed superior sensitivity and fewer false positives compared to previous frameworks. Note that the higher sensitivity obtained, the more false positive pulmonary nodules resulted. Our architecture was shown to tackle the problems of high false positive rate caused by improving recall.

In the lung cancer screening step, radiologists will generally take a long time to read and analyze CT scans to make the correct clinical interpretation. But there are many factors making experienced radiologist prone to misdiagnosis, such as multi-sequence /multi-modality of images, the tiny size

3. In the abbreviation of the table: Calc. represents calcified nodules; Pleu. represents nodules on the pleura; 3-6, 6-10, 10-30 represents the longest diameter of solid nodules (mm); Mass. represents the case of solid nodules' longest diameter larger than 30 mm; p.ggn denotes pure GGN and m.ggn denotes mix ggn, or sub-solid nodules.

Table 3: Ablation study with the RCNN series, YOLO and SSD series on our chest CT dataset. The entries SSD300 used the input image resolution as the  $300 \times 300$  with the backbone of ResNeXt, and we use the SSD512 without bells and whistles as the baseline. FP rate represents the ratio of false positive (FP) to true positive (TP). Detailed information on the eight classes can be found in footnote 3.

Method	CPM	FP rate	Calc.	Pleu.	3-6	6-10	10-30	Mass	p.ggn	m.ggn
RCNN (Ren et al., 2015)	0.464	1.30	83.8	55.6	77.4	90.5	84.4	77.8	83.9	89.7
RCNN (He et al., 2016)	0.517	1.17	89.1	62.9	81.3	94.6	93.8	100	83.2	91.2
RCNN (Xie et al., 2017)	0.538	0.99	86.9	62.4	78.9	91.9	93.8	100	86.1	92.6
SSD300 (Liu et al., 2016)	0.492	1.28	91.0	68.4	84.7	90.5	93.8	100	86.9	92.6
SSD512 (Liu et al., 2016)	0.533	1.21	91.0	63.2	81.0	91.9	96.9	100	78.8	85.3
YOLO (Redmon et al., 2016)	0.499	1.30	90.6	65.2	81.4	91.5	93.8	100	86.3	90.9
SSD300(ResNeXt)	0.546	1.15	92.2	65.0	84.0	85.1	93.8	100	73.7	70.6
SSD512(ResNeXt)	0.555	1.35	92.2	65.5	84.2	87.8	96.9	100	85.4	85.3
3DCNN(Wang et al., 2018a)	0.700	1.59	91.3	60.3	80.5	91.9	93.8	100	85.0	91.2
GA-SSD512(ours)	<b>0.733</b>	<b>0.89</b>	90.7	65.0	82.7	93.2	93.8	100	<b>94.2</b>	<b>97.1</b>

and low density of some lesions (such as GGN) that signal early lung cancer, heavy workload, and the repetitive nature of the job. Our proposed CNN-based system for pulmonary nodules detection achieve state-of-the-art performance with low false positives rate. Moreover, our proposed model takes only nearly 30s to detect pulmonary nodules, and it still has the potential to further speed up the detection process when more computing resources are available.

## References

- Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.
- Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries: Global cancer statistics 2018. *CA: A Cancer Journal for Clinicians*, 09 2018.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017a.
- Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, volume 2, page 3, 2017b.

- Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Rydhwana Hossain, Carol C Wu, Patricia M de Groot, Brett W Carter, Matthew D Gilman, and Gerald F Abbott. Missed lung cancer. *Radiologic Clinics of North America*, 2018.
- Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- Tsung Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 936–944, 2016.
- Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *arXiv preprint arXiv:1809.02165*, 2018.
- Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- Qianhui Luo, Huifang Ma, Yue Wang, Li Tang, and Rong Xiong. 3d-ssd: Learning hierarchical features from rgb-d images for amodal 3d object detection. *arXiv preprint arXiv:1711.00238*, 2017.
- David J Manning, SC Ethell, and Tim Donovan. Detection or decision errors? missed lung cancer from the posteroanterior chest radiograph. *The British journal of radiology*, 77(915):231–235, 2004.
- Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Matthias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- A. A. Setio, F Ciompi, G Litjens, P Gerke, C Jacobs, Riel S Van, Wille M Winkler, M Naqibullah, C Sanchez, and Ginneken B Van. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, 35(5): 1160–1169, 2016.
- Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42: 1–13, 2017.
- Miller Siegel and Jemal. Cancer statistics, 2018. *CA: a cancer journal for clinicians*, 68(1):7–30, 2018.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. *arXiv preprint arXiv:1806.05372*, 2018.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. URL <http://arxiv.org/abs/1409.4842>.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- Bin Wang, Guojun Qi, Sheng Tang, Liheng Zhang, Lixi deng, and Yongdong Zhang. Automated pulmonary nodule detection: High sensitivity with few candidates. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 759–767, September 2018a.
- Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018b.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5987–5995. IEEE, 2017.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *arXiv preprint arXiv:1512.05193*, 2015.

Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Int. Conf. on Computer Vision*, volume 6, 2017.

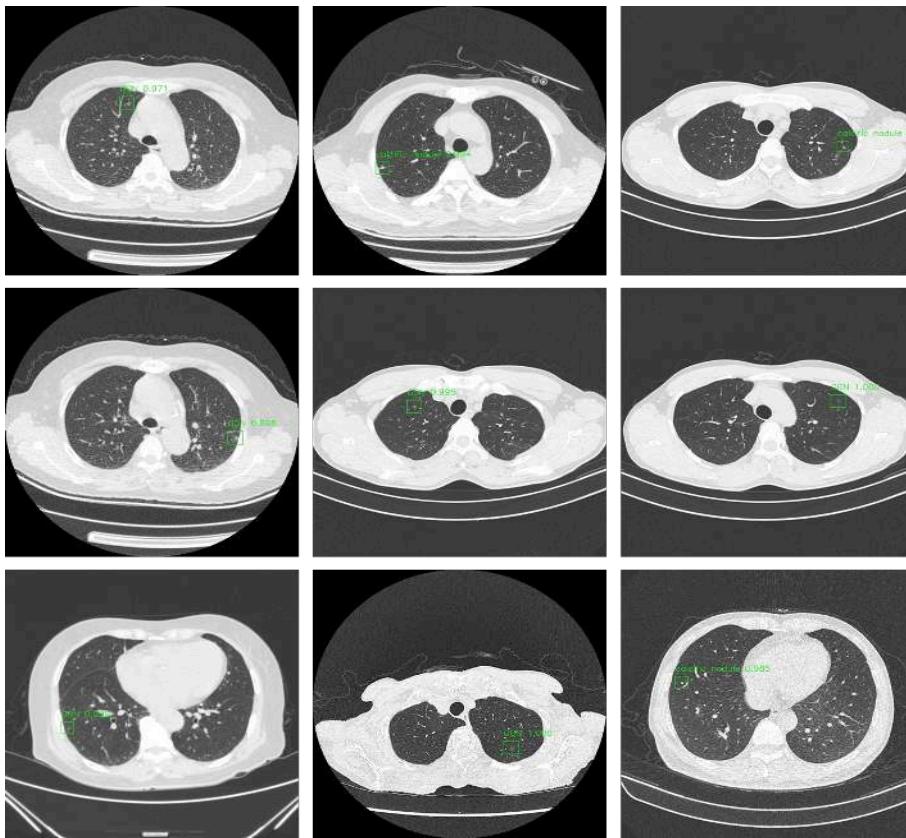
**Appendix A. Sample results of the detection model**

Figure 4: Results of true positives on nine cases. These nodules including the ones with small size are difficult to identify but are detected by our model.



Figure 5: Results of false positives on three cases. These false positives have similar appearances with the nodules and are easily detected as abnormalities.

# A novel segmentation framework for uveal melanoma in magnetic resonance imaging based on class activation maps

**Huu-Giao Nguyen<sup>1,2,3</sup>**

HUU.NGUYEN@ARTORG.UNIBE.CH

<sup>1</sup> Proton Therapy Center, Paul Scherrer Institut, ETH Domain, Villigen, Switzerland

<sup>2</sup> Ophthalmic Technology Lab., ARTORG Center, University of Bern, Switzerland

<sup>3</sup> Radiology Department, Lausanne University Hospital (CHUV), Switzerland

**Alessia Pica<sup>1</sup> and Jan Hrbacek<sup>1</sup> and Damien C. Weber<sup>1,4</sup>**

FIRSTNAME.LASTNAME@PSI.CH

<sup>4</sup> Radiation Oncology Department, Inselspital, University of Bern, Switzerland

**Francesco La Rosa<sup>5</sup>**

FRANCESCO.LAROSA@EPFL.CH

<sup>5</sup> Signal Processing Lab., Ecole Polytechnique Fédérale de Lausanne, Switzerland

**Ann Schalenbourg<sup>6</sup>**

ANN.SCHALENBOURG@FA2.CH

<sup>6</sup> Adult Ocular Oncology Unit, Jules-Gonin Eye hospital, Lausanne, Switzerland

**Raphael Sznitman<sup>2</sup>**

RAPHAEL.SZNITMAN@ARTORG.UNIBE.CH

**Meritxell Bach Cuadra<sup>3,5,7</sup>**

MERITXELL.BACHCUADRA@UNIL.CH

<sup>7</sup> Medical Image Analysis Laboratory, CIBM, University of Lausanne, Switzerland

## Abstract

An automatic and accurate eye tumor segmentation from Magnetic Resonance images (MRI) could have a great clinical contribution for the purpose of diagnosis and treatment planning of intra-ocular cancer. For instance, the characterization of uveal melanoma (UM) tumors would allow the integration of 3D information for the radiotherapy and would also support further radiomics studies. In this work, we tackle two major challenges of UM segmentation: 1) the high heterogeneity of tumor characterization in respect to location, size and appearance and, 2) the difficulty in obtaining ground-truth delineations of medical experts for training. We propose a thorough segmentation pipeline consisting of a combination of two Convolutional Neural Networks (CNN). First, we consider the class activation maps (CAM) output from a Resnet classification model and the combination of Dense Conditional Random Field (CRF) with a prior information of sclera and lens from an Active Shape Model (ASM) to automatically extract the tumor location for all MRIs. Then, these immediate results will be inputted into a 2D-Unet CNN whereby using four encoder and decoder layers to produce the tumor segmentation. A clinical data set of 1.5T T1-w and T2-w images of 28 healthy eyes and 24 UM patients is used for validation. We show experimentally in two different MRI sequences that our weakly 2D-Unet approach outperforms previous state-of-the-art methods for tumor segmentation and that it achieves equivalent accuracy as when manual labels are used for training. These results are promising for further large-scale analysis and for introducing 3D ocular tumor information in the therapy planning.

**Keywords:** Activation map, CAM, Unet, tumor segmentation, Uveal melanoma

## 1. Introduction

UM is the most common primary intraocular malignancy in the white adult population, making up 79-88% of primary intraocular cancers (Singh et al., 2014; Lemke et al., 1999). Several 2-

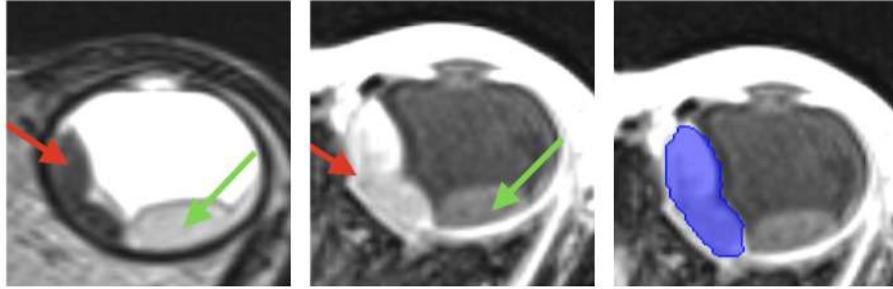


Figure 1: Example of UM in MRI: (left) T2-w; (center) T1-w; (right) manual tumor segmentation. Red & green arrows indicate the tumor & retinal detachment respectively.

dimensional and 3-dimensional imaging modalities such as 2D Fundus imaging and 3D computed tomography are needed to properly characterize the tumor, its growth and for any follow-up care. Recently, 3D MRI is raising interest in the treatment of ocular tumors (de Graaf et al., 2014; Tartaglione et al., 2014). Thanks to its high spatial resolutions and high intrinsic contrast, 3D MRI allows for clear overall improved discrimination between anatomical structures and different pathological regions (Tartaglione et al., 2014) (see Fig. 1). An automatic extraction of quantitative and reliable information of ocular tumors in MR images, i.e the location, size, texture, morphology and distribution of pathological tissues, would be a breakthrough in current diagnosis, follow-up and therapy planning procedures. Ultimately, having 3D patient-specific eye model provide an optimal solution for radiation therapy in the framework of personalized medicine in order to plan and deliver very conformal radiation dose to the tumor while minimizing irradiation of critical structures (Beenakker et al., 2015).

Few automated methods have been tailored for ocular tumors segmentation from MRI. First attempts were dedicated to the segmentation of retinoblastoma in children. Two deep learning techniques were proposed, a 3D-Unet (Nguyen et al., 2018a) and a 3D-CNN (Ciller et al., 2017), based on rather small datasets of 16 and 32 retinoblastoma eyes. The tumor segmentation performance reported in those pioneer works was, however, relatively low, with an average Dice similarity coefficient (DSC) measurement of around 62%. Recently, UM tumors have been tackled in (Hassan et al., 2018), based on image registration and threshold of MRI, though only four cases were qualitatively evaluated. One of the major limitations of these approaches, affecting specifically supervised techniques, is the lack of manual delineations. Actually, we think the above 3D deep learning architectures were highly limited by the low number of training samples available. Unfortunately, having such input labels is very tedious, time consuming and not easily available in practice.

Weakly supervised methods based on CAM for the segmentation of pathological tissues have recently received a great attention, e.g. pulmonary nodules in CT (Feng et al., 2017) or diabetic retinopathy lesions in retinal fundus images (Gondal et al., 2017). Here, our first aim is to present an ocular tumor segmentation framework without the need of manual annotations for training. To this end, we propose an end-to-end tumor segmentation framework with two CNNs for 2D images extracted from 3D volume MRI. Our approach is based on the estimation of CAMs from a CNN architecture that classifies whether there is a tumor or not in the image. Afterwards, we refine the CAMs by combining an ASM segmentation of the eye structures (Nguyen et al., 2018b) with a dense

Table 1: MR imaging acquisition parameters at 1.5T with a surface coil.

	Repetition time(ms)	Echo time (ms)	Flip Angle	Voxel size ( $mm^3$ )	FOV (Voxels)	Healthy	UM
T1-VIBE	6.55	2.39	12°	0.5x0.5x0.5	256x256x80	28 eyes	24 eyes
T2-SPACE	1400	185	150°	0.5x0.5x0.5 and 0.82x0.82x0.8	256x256x80	25 eyes	22 eyes

CRF to maximize label agreement between similar pixels in images. Finally, we use these refined CAMs as input training data for a 2D-Unet segmentation (Ronneberger et al., 2015). The proposed framework is cheaper in training data (only sclera segmentations are needed for the ASM) and outperforms in segmentation compared existing deep learning approaches (Nguyen et al., 2018a; Rosa et al., 2018).

A second major contribution of this work is the quantitative evaluation of several 2D and 3D architectures for the UM segmentation. To the best of our knowledge this is the first study reporting automated segmentation accuracy for such ocular tumor. Our proposed segmentation technique will be compared with previous related work: 1) our previous work tailored for retinoblastoma tumors in children and based on a 3D-Unet (Nguyen et al., 2018a), with a 2D-Unet using manual labels as training from an expert, and 2) a cascade of two 3D patch-wise CNNs used for lesion segmentation in Multiple Sclerosis (Rosa et al., 2018).

## 2. Dataset

MR acquisitions were performed by a 1.5T Siemens scanner with surface coil for both T1w and T2w contrasts at the Paul Scherrer Institute. A set of 16 healthy volunteers (mean age  $29 \pm 5.4$  y.o., range [23 – 46] years) and 24 UM patients (mean age  $63 \pm 14$  y.o., range [36 – 74] years) was considered. The cohort median eye size was 24.4mm of diameter (range, 22.1-26.5). Tab. 1 shows the different parameters used for the MRI acquisition protocol. The study was approved by the Ethics Committee of the involved institutions and all subjects (anonymized and de-identified) provided written informed consent prior to participation.

Images were pre-processed as follows. First, an anisotropic diffusion filtering (Perona and Malik, 1990) was applied to reduce noise without removing significant image content. Second, we applied the N4 algorithm (Tustison et al., 2010) to correct for bias field variations and performed histogram-based intensity normalization (Nyul et al., 2000) for an intensity normalisation. Finally, in order to improve the performance in segmentation and computation time, the whole MRI was cropped using a volume of interest of 64x64x64 voxels centered in the eye.

Manual delineations were done by radiation oncologist expert for 16 UM patients and all healthy eyes using Velocity software(Varian Medical System, Palo Alto, CA). First, segmentation for sclera, lens and tumor was done individually through intensity threshold. Second, manual editing was performed to refine borders and remove outlier regions.

## 3. Proposed segmentation framework

The proposed framework is over-viewed in Fig. 2. It mainly consists of the concatenation of a 2D ResNet model (He et al., 2016) to classify MRI slices (with or without tumor) that combined with

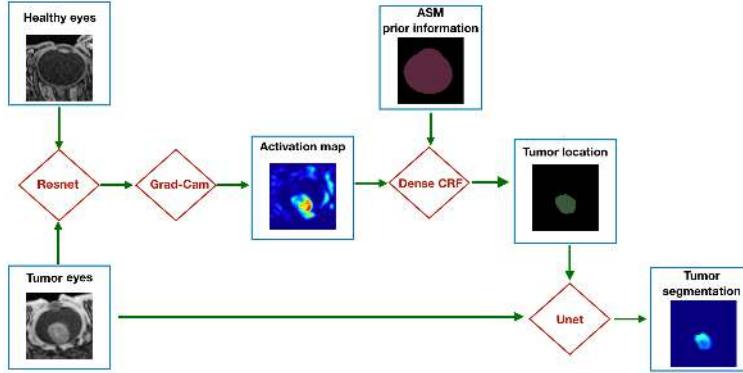


Figure 2: Main pipeline of our approach proposed.

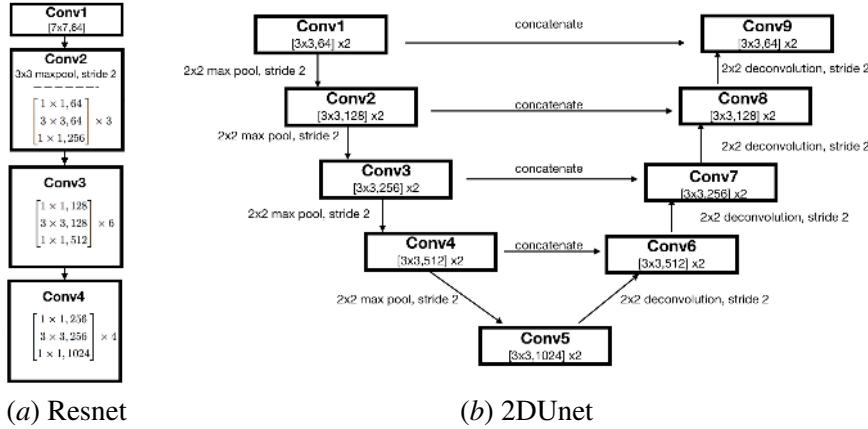


Figure 3: Neural network architectures used.

the ASM (Nguyen et al., 2018b) and a CRF (Krähenbühl and Koltun, 2011) will provide the initial labels for training a 2D-Unet (Ronneberger et al., 2015) model to segment tumor.

**ResNet classification.** In this work, we used the ResNet model (He et al., 2016) for classification of 2D input images with the score presence or absence of tumor. ResNet has the advantage of avoiding the degradation problem of deep CNN, which occurs when the accuracy gets saturated and rapidly degrades as a result of an increasing network depth. The ResNet replaces a direct mapping of input  $x$  to its score  $y$  with a function  $F(x)$  by a residual function using  $F(x) + x$ , where  $F(x)$  and  $x$  represents the stacked non-linear layers and the identity function respectively. The architecture of our ResNet is in Fig. 3(a).

**Tumor location by CAM.** Considering a CNN-based classification, each layer retains detailed spatial information of object and its characterization used by network to identify the category. CAMs (Zhou et al., 2016) produce such class-discriminative localization using a linear combination of  $f_k(i, j)$  represent the activation of unit  $k$  in the last convolutional layer at spatial location  $(i, j)$  and the weight  $w_k^c$  corresponding to class  $c$  for unit  $k$ :

$$M^c(i, j) = \sum_k w_k^c f_k(i, j) \quad (1)$$

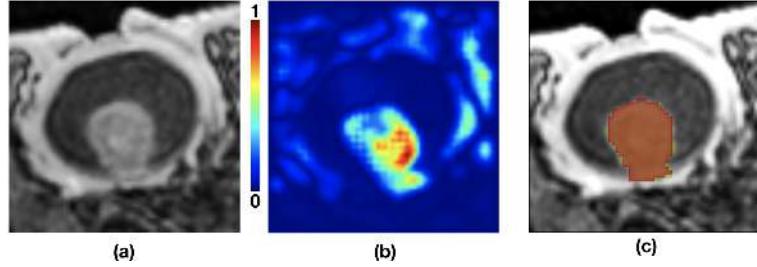


Figure 4: Example of tumor location: (a) original image; (b) grad-cam result; (c) ASM constrain for sclera (in red), DenseCRF result (in green - Dice overlap 95%).

As a generalization of CAM, Grad-CAM heat map (Selvaraju et al., 2017) is constructed from the liner combination of the importance weights  $\alpha_k^c$  and feature maps  $A$  of a convolutional layer with respect to the gradient of the score  $y^c$  of class  $c$ :

$$L^c = \text{ReLU} \left( \sum_k \alpha_k^c A_k \right) \quad (2)$$

**Refinement.** We apply a dense CRF (Krähenbühl and Koltun, 2011) to maximize label agreement between similar pixels. The Dense CRF incorporates unary potentials of individual pixels and pair-wise potentials (in terms of appearance and smoothness) on neighboring pixels to provide more homogeneous regions. Considering as input image the 2D MRI slice  $I$  (either T1w or T2w) and a probability map  $P$  provided by the Grad-CAM, the unary potential is defined to be the negative log-likelihood  $\psi_u(z_i) = -\log P(z_i|I)$ , where  $z_i$  the predicted label of voxel  $i$ . The pair-wise potential has the form  $\psi_p(z_i, z_j) = \mu(z_i, z_j)k(f_i, f_j)$ , where  $\mu$  is a label compatibility function and  $k(f_i, f_j)$  is characterized by integrating two Gaussian kernels of appearance (first term) and smoothness (second term), as follows:

$$k(f_i, f_j) = w_1 \exp \left( -\frac{|p_i - p_j|^2}{2\theta_1^2} - \frac{|I_i - I_j|^2}{2\theta_2^2} \right) + w_2 \exp \left( -\frac{|p_i - p_j|^2}{2\theta_3^2} \right), \quad (3)$$

where  $p_i$  are pixel locations,  $I_i$  are pixel intensities,  $w_l$  are weight factor between the two terms, and the  $\theta$ 's are tunable parameters of the Gaussian kernels. The Gibbs energy of CRF model is then given by  $\sum(\psi_u(z_i), \psi_p(z_i, z_j))$ . Here, we apply the inference of Dense CRF with different iterative numbers  $\{5, 20, 50\}$  where the mean field approximation is computed by minimizing the KL-divergence while constraining the distributions.

Finally, prior information about the healthy structures such as the sclera and lens was used as tumor location constraint. Our previous work (Nguyen et al., 2018b) evaluated the DSC values of the sclera ( $94.5\% \pm 1.6$ ) and lens ( $88.3\% \pm 2.8$ ) on the same data set. The ASM segmentation is used to constrain the result of the CRF as shown in Fig. 4.

**UNet.** Similar to the original UNet method (Ronneberger et al., 2015), we consider an encoder and decoder network that takes as input 2D image with tumor and label output of Grad-cam. Each encoding pathway contains 4 layers that effectively changes the feature dimension (i.e. 64, 128, 256, 512, 1024 - Fig. 3(b)). The same architecture accounts for the decoding pathway. In each case, two

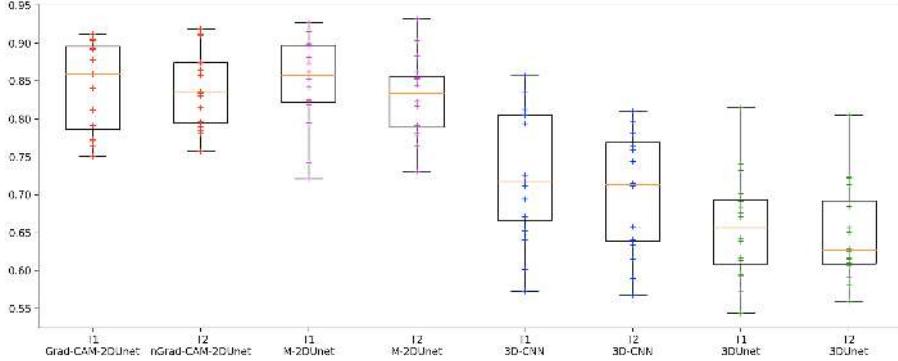


Figure 5: Boxplot on DSC of tumors segmentation on T1-w and T2-w MR images.

Wilcoxon test compared with	M-2DUnet	3D-CNN	3D-Unet
<i>p_value</i> on T1	0.679	0.00320	0.000437
<i>p_value</i> on T2	0.147	0.00097	0.000436

Table 2: Wilcoxon signed rank test on DSC comparing our method with the other strategies.

convolution layers 3x3 are used with rectified linear unit (Relu) operations and with zero padding. Between two layers in the encoder pathway, 2x2 max pooling with strides of two in each dimension are used. In the decoder pathway, a 2x2 deconvolution layer with strides of two is firstly used. Concatenation is performed to connect the output tensors of two layers of the encoder and decoder pathways at same level. To train our network, we used the Adam optimizer and the binary cross entropy loss function. Softmax is used to extract probability maps for each class. Data augmentation including rotation, shift as well as elastic deformation was applied (Simard et al., 2003).

#### 4. Quantitative evaluation

We computed the DSC value of the predicted output as compared to the manual segmentation for the quantitative evaluation of all the automated techniques. For the 16 patients with manual segmentation, we used a leave-one-out cross-validations strategy, i.e., iteratively chose one eye as a test case, two other random eyes as validation cases while the remaining subjects are used as the training set. Moreover, to show the advantage of the proposed weakly learning 8 additional patients without manual segmentation are also included into the training set. The average number of 2D slices (containing the tumor) extracted from 3D volume of patient's eyes is 45 (range [25-60]), overall is 925 images.

Resnet binary classification model construction is trained including also tumor-free eyes. In this stage, 1915 2D slices extracted from 28 healthy eyes are also added into training set, i.e our training set have 2840 images of healthy and pathological eyes. CAMs are estimated based on all 2D images of 24 UM patient. For 2D-Unet, depending on the patients leaved out for test and validation set, around 850 2D images with tumors were selected for training.

Our framework (Grad-CAM-2DUnet) is evaluated in comparison with three baseline deep learning architectures. First, the same 2D-Unet architecture included in our framework will be used trained with the expert manual delineations instead of using the refined activation maps (M-2DUnet).

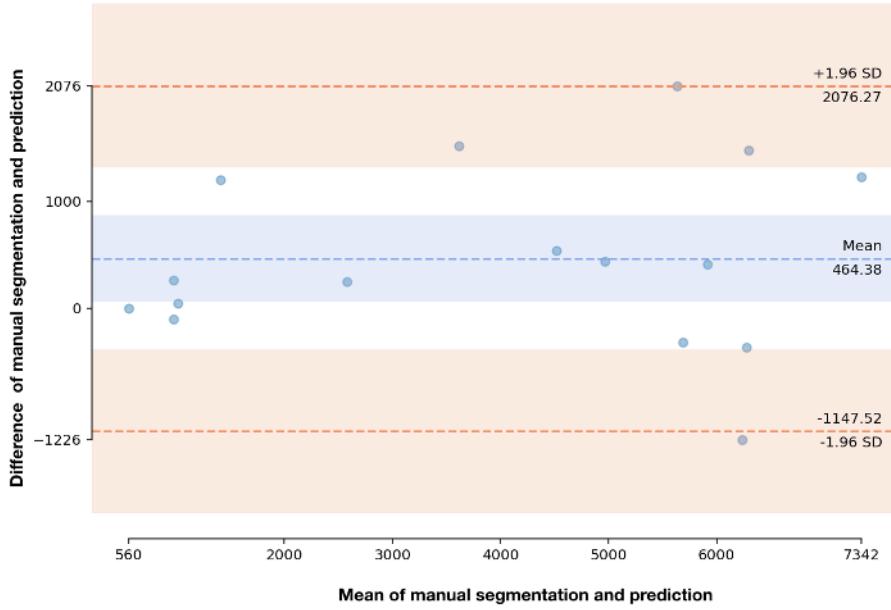


Figure 6: The Bland-Altman plot of differences (number of volume voxels) between Grad-CAM-2DUnet method's result and manual segmentation.

Second, our previous 3D-Unet ([Nguyen et al., 2018a](#)) tested on retinoblastoma patients (3D-Unet). It is composed of 4 layers of encoder and decoder pathway with different feature sizes (i. e., 32, 64, 128, 256, 512); 3x3x3 convolution with PReLU and 2x2x2 max pooling. Third, a cascade of two 3D patch-wise convolutional neural networks ([Valverde et al., 2017](#); [Rosa et al., 2018](#)) (3D-CNN) that reported high accuracy in segmenting white matter lesions. It is composed of with 4 convolutional layers ([3x3x3,32]x2; [3x3x3,64]x2); patch-size is 11x11x11 (input images interpolated to 256x256x256).

[Fig.5](#) shows the boxplot on DSC of four tumor segmentation methods for both T1-w and T2-w sequences, where 3DUnet with  $65.8 \pm 6.8$  ( $64.9 \pm 6.3$ ) and 3DCNN with  $72.6 \pm 8.2$  ( $70.5 \pm 7.5$ ) perform in average 10% worst than the 2D-Unet strategies. This can be explained by the increased training set available in 2D as compared to the few training data in 3D. Thus, despite image acquisition is done in 3D and with a very nice isotropic spatial resolution, 2D approaches perform better. Let us note that differences in DSC were statistically significant (Wilcoxon signed rank test,  $p < 0.005$ ) when comparing Grad-CAM-2DUnet with 3D-UNET and 3D-CNN in both T1w and T2w scenarios ([Tab.2](#)).

Our weakly supervised framework Grad-CAM-2DUnet with average Dice of  $84.5 \pm 5.6$  for T1-w and  $83.9 \pm 4.9$  for T2-w performs similarly to the M-2DUnet (using muanual segmentations for training) with  $84.8 \pm 5.7$  and  $82.9 \pm 5.2$  for T1-w and T2-w, respectively. No statistical differences were found neither for T1-w nor T2-w. The mean False positive and True positive fractions are 0.02 and 0.82 respectively when we compared our Grad-CAM-2DUnet prediction with manual segmentation. [Fig.6](#) shows a Bland-Altman difference plot of the 3D volume comparison of manual segmentation and our prediction. This result shows the relevance of our solution for replacing the

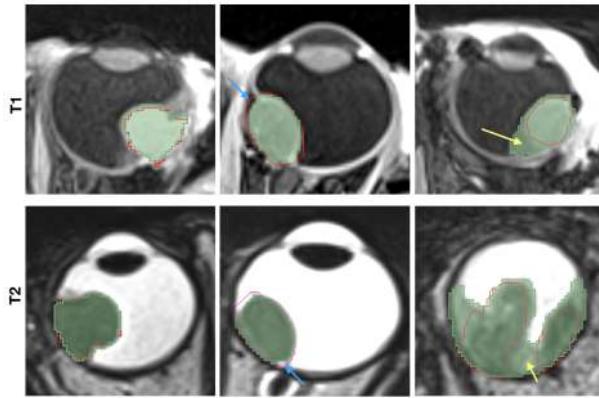


Figure 7: Qualitative results of our tumor prediction (in green) & manual segmentation (in red) on T1w and T2w: (left) high overlap; (center) automated segmentation fixed better the intensity contour (blue arrows); (right) low accuracy: our prediction cannot separate tumor and retinal detachment (yellow arrows).

costly manual annotations by free refined activation maps. Fig. 7 shows qualitative result of the tumor segmentation with our proposed approach.

## 5. Conclusion

In this paper, we introduced an automatic and effective deep learning based approach that allows a quantitative image analysis of eye tumor tissues in adults that could further support clinicians to tailor the radiation therapy to the UM in eye tumor patients. The proposed approach takes advantage of CAMs combined with conditional random field and active shape models to provide an end-to-end segmentation without need of tumor annotations of medical experts. The paper also provides an evaluation of several 2D and 3D deep learning strategies for the UM segmentation. To our knowledge, this is the first set of techniques that have been proposed for the segmentation of UM, reporting very high accuracy in average. Our study, based on a 3D high-resolution dataset of 24 tumors, demonstrates that the best strategies for tumor segmentation make use of 2D slices instead of 3D whole volumes, that is including more data for training. Our weakly supervised framework provides a solid reliable computer-aided tool to further large-scale evaluation of ocular tumors based on MR imaging features for an enhancing a shift towards non-invasive clinical procedures.

## Acknowledgments

This work is funded by the Swiss Cancer Research foundation (grant no. GAP-CRG-201602) and is supported by the Center of Biomedical Imaging of Geneva-Lausanne Universities and EPFL, the Fondation Leenaards and Fondation Louis-Jeantet. FLR is funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie project TRABIT (agreement No 765148).

## References

- J. Beenakker, D. Shamonin, A. Webb, G. Luyten, and B. Stoel. Automated retinal topographic maps measured with magnetic resonance imaging. *Investig. Ophthalmol. Vis. Sci.*, 56:1033–1039, 2015.
- C. Ciller, S. De Zanet, K. Kamnitsas, P. Maeder, B. Glocke, F. Munier, D. Rueckert, J-P. Thiran, M. Bach Cuadra, and R. Sznitman. Multi-channel mri segmentation of eye structures and tumors using patient-specific features. *PLoS ONE*, 12(3), 2017.
- P. de Graaf, S. Görliche, F. Rodjan, P. Galluzzi, P. Maeder, J. Castelijns, and H. Brisse. Guidelines for imaging retinoblastoma: imaging principles and mri standardization. *Pediatric radiology*, pages 2–14, 2014.
- X. Feng, J. Yang, A. Laine, and E. Angelini. Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules. *MICCAI*, page 568–576, 2017.
- W. Gondal, J. Köhler, R. Grzeszick, G. Fink, and M. Hirsch. Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. *arXiv preprint arXiv:1706.09634*, 2017.
- M. Hassan, D. Shamonin, R. Shahzad, A. Webb, B. Stoel, and J-W. Beenakker. Automated analysis of eye tumor mr-images for an improved treatment determination. *ISMRM*, 2018.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016.
- P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. in Neural Information Processing Systems*, pages 109–117, 2011.
- A. Lemke, N. Hosten, N. Bornfeld, N. Bechrakis, A. Schüler, M. Richter, C. Stroszczynski, and R. Felix. Uveal melanoma: correlation of histopathologic and radiologic findings by using thin-section mr imaging with a surface coil. *Radiology*, 210(3):775–783, 1999.
- H-G. Nguyen, A. Pica, P. Maeder, A. Schalenbourg, M. Peroni, J. Hrbacek, DC. Weber, M. Bach Cuadra, and R. Sznitman. Ocular structures segmentation from multi-sequences mri using 3d unet with fully connected crfs. *Comput. Path. & Opht. Med. Image Analysis*, pages 167–175, 2018a.
- H-G. Nguyen, R. Sznitman, P. Maeder, A. Schalenbourg, M. Peroni, J. Hrbacek, DC. Weber, A. Pica, and M. Bach Cuadra. Personalized anatomic eye model from t1-weighted vibe mr imaging of patients with uveal melanoma. *Journal of Radiation Oncology, Biology, Physics*, 2018b.
- L.G. Nyul, J.K. Udupa, and Xuan Zhang. New variants of a method of mri scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–50, 2000.
- P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 12(7):629–639, 1990.
- O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 9351:234–241, 2015.

- F. La Rosa, M. Fartaria, T. Kober, J. Richiardi, C. Granziera, J-P. Thiran, and M. Bach Cuadra. Shallow vs deep learning architectures for white matter lesion segmentation in the early stages of multiple sclerosis. *MICCAI workshop*, 2018.
- R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *In arXiv:1610.02391v3*, 2017.
- P.Y. Simard, D. Steinkraus, and J.C. Platt. Best practices for convolutional neural networks applied to visual document analysis. *ICDAR*, pages 958–963, 2003.
- A. Singh, L. Bergman, and S. Seregard. Uveal melanoma: epidemiologic aspects. *Clini. Opht. Onc.*, page 75–87, 2014.
- T. Tartaglione, M. Pagliara, M. Sciandra, C. Caputo, R. Calandrelli, G. Fabrizi, S. Gaudino, M. Blasi, and C. Colosimo. Uveal melanoma: evaluation of extrascleral extension using thin-section mr of the eye with surface coils. *Radiol Med.*, 119(10):775–783, 2014.
- N. Tustison, B. Avants, P. Cook, Y. Zheng, A. Egan, P. Yushkevich, and J. Gee. N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010.
- S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J-C. Vilanova, L. Ramió-Torrentà, A. Rovira, A. Oliver, and X. Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. *NeuroImage*, 2017.
- B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *CVPR*, pages 2921–2929, 2016.

# High-quality segmentation of low quality cardiac MR images using k-space artefact correction

**Ilkay Oksuz<sup>1</sup>**

ILKAY.OKSUZ@KCL.AC.UK

**James Clough<sup>1</sup>**

JAMES.CLOUGH@KCL.AC.UK

**Wenjia Bai<sup>2</sup>**

W.BAI@IMPERIAL.AC.UK

**Bram Ruijsink<sup>1</sup>**

JACOBUS.RUIJSINK@KCL.AC.UK

**Esther Puyol-Antón<sup>1</sup>**

ESTHER.PUYOLANTON@KCL.AC.UK

**Gastao Cruz<sup>1</sup>**

GASTAO.CRUZ@KCL.AC.UK

**Claudia Prieto<sup>1</sup>**

CLAUDIA.PRIETO@KCL.AC.UK

**Andrew P. King<sup>1</sup>**

ANDREW.KING@KCL.AC.UK

**Julia A. Schnabel<sup>1</sup>**

JULIA.SCHNABEL@KCL.AC.UK

<sup>1</sup> School of Biomedical Engineering & Imaging Sciences, King’s College London, UK

<sup>2</sup> Data Science Institute and Department of Medicine, Imperial College London, UK

## Abstract

Deep learning methods have shown great success in segmenting the anatomical and pathological structures in medical images. This success is closely bounded with the quality of the images in the dataset that are being segmented. A commonly overlooked issue in the medical image analysis community is the vast amount of clinical images that have severe image artefacts. In this paper, we discuss the implications of image artefacts on cardiac MR segmentation and compare a variety of approaches for motion artefact correction with our proposed method Automap-GAN. Our method is based on the recently developed Automap reconstruction method, which directly reconstructs high quality MR images from k-space using deep learning. We propose to use a loss function that combines mean square error with structural similarity index to robustly segment poor-quality images. We train the reconstruction network to automatically correct for motion-related artefacts using synthetically corrupted CMR k-space data and uncorrected reconstructed images. In the experiments, we apply the proposed method to correct for motion artefacts on a large dataset of 1,400 subjects to improve image quality. The improvement of image quality is quantitatively assessed using segmentation accuracy as a metric. The segmentation is improved from 0.63 to 0.72 dice overlap after artefact correction. We quantitatively compare our method with a variety of techniques for recovering image quality to showcase the influence on segmentation. In addition, we qualitatively evaluate the proposed technique using k-space data containing real motion artefacts.

**Keywords:** Cardiac MR Segmentation, Image Quality, Image Artefacts, Image Artefact Correction, Deep Learning, UK Biobank, Automap

## 1. Introduction

Image segmentation is an extensively investigated problem in medical imaging, for which deep learning methods have demonstrated considerable success (Litjens et al., 2017). In general, neural network architectures are trained using controlled databases, but their performance does not always translate to clinical data in practice. Variability of image acquisition protocols, the presence of

pathology (Shao et al., 2018) and image artefacts (Oksuz et al., 2018b) can all cause low segmentation accuracy. In particular for cardiac magnetic resonance (CMR) images, which can contain a variety of different imaging artefacts (Ferreira et al., 2013), it is often difficult to segment structures and extract cardiac indices using trained neural networks on data ‘in the wild’, which hinders the translation of deep learning models into clinical practice. In this work, we first highlight the low segmentation accuracy when the original image is of low quality. We then propose a pipeline to improve the image quality, which translates into improved accuracy in the subsequent segmentation task. We compare a wide range of methods that can address the problem of image quality in CMR imaging. For poor quality CMR images, ground truth segmentations are usually not available, and traditionally, these low quality images are excluded from further analysis, leading to the need to recall patients to reacquire images. In the context of cohort studies, excluding individual subjects diminishes the research value of the evaluation.

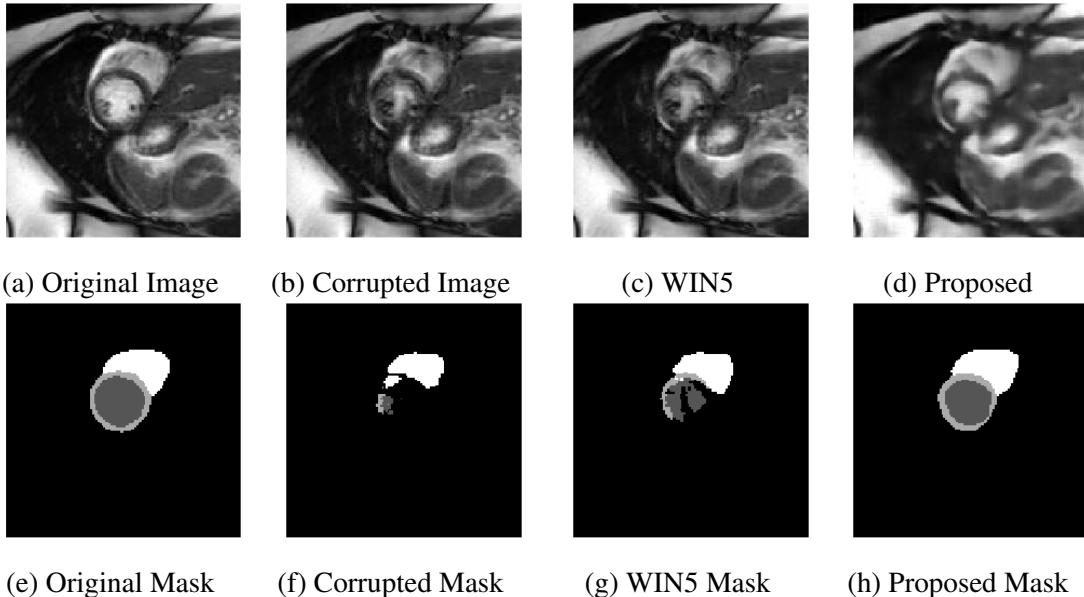


Figure 1: Image quality is a prerequisite for high segmentation accuracy. The segmentation accuracy of the state-of-the-art network (Bai et al., 2018) drops with degradation of image quality (e,f). Our proposed method is able to generate high quality images (d), which improves the segmentation (h) compared to a state-of-the-art image denoising technique (Liu and Fang, 2017) (c,g)

Common motion artefacts in CMR manifest as temporal and/or spatial blurring of the images, which makes subsequent processing difficult (Ferreira et al., 2013). Examples of a good quality image and a synthetically motion-corrupted image are shown in Fig.1a and 1b for a short-axis view CINE CMR scan. Applying a state-of-the-art CMR segmentation algorithm (Bai et al., 2018) on these images, one can see that the method is capable of segmenting the left ventricle, myocardium and right ventricle with considerable success when the original image is of good quality (Fig.1e). However, when the image quality is reduced using a synthetic corruption strategy proposed in (Oksuz et al., 2018b), the segmentation performance diminishes (Fig.1f). To improve the image quality

we use an end-to-end image artefact correction algorithm that uses k-space input. Our algorithm outputs high quality images (Fig.1d) and correspondingly improved segmentations (Fig.1h) compared to state-of-the-art image denoising techniques (Liu and Fang, 2017) (Fig.1c and g).

Our approach is based on automatically correcting for artefacts during the reconstruction process based on our previous work (Oksuz et al., 2018a). We use a deep neural network for correcting artefacts and evaluate our method on a synthetic dataset of 1400 2D+time CMR images from the UK Biobank (Petersen et al., 2015). We also evaluate the performance on real artefact cases to demonstrate the performance of our method. There are three major contributions of this work. First, we illustrate the implications of low image quality on CMR segmentation in an extensive study. Second we propose to use motion artefact correction directly from k-space using a novel loss function. Finally, we provide a thorough investigation on the influence of using different artefact correction mechanisms on CMR segmentation.

## 2. Background

Deep learning techniques have been utilized for segmentation problems with high success (Litjens et al., 2017). However the influence of variability in acquisition protocols, pathology and image artefacts is an often overlooked problem. In a recent work Shao et al. (2018) showed the shortcomings of deep learning in the presence of pathology for brain MRI segmentation with an emphasis on the selection of training data.

Estimating high quality images from corrupted (or under-sampled) k-space has been a more investigated subject in the literature (Han and Ye, 2018). The problem can be addressed either in the k-space domain or the image domain. One choice is to correct the k-space before applying the inverse Fourier transform (IFT) as proposed by Han and Ye (2018). A more common approach is to use the IFT on k-space and learn a mapping between the corrupted reconstructed images and good quality images. To this end a variety of image denoising techniques can be utilized such as autoencoders (Xie et al., 2012), residual learning networks (Zhang et al., 2017) or wide networks (Liu and Fang, 2017).

In the context of CMR artefact correction, early works focused on changes in acquisition schemes (Saremi et al., 2008) and analytical methods for motion artefact reduction (Kim et al., 2008). For automatic correction of the CMR, Lötjönen et al. (2005) used short-axis and long-axis images to optimize the locations of the slices using mutual information as a similarity measure. More recently deep learning methods have been utilized to accelerate MR image acquisition by using undersampling in k-space. Schlemper et al. (2018a) proposed to use a deep cascaded network to generate high quality images. In a more recent work the authors proposed to use deep latent representations for myocardial segmentation from low quality images (Schlemper et al., 2018b). Hauptmann et al. (2019) proposed to use a residual U-net to reduce aliasing artefacts due to undersampling with the purpose of accelerating image acquisition. Our work differs from these works, which rely on a relatively small number of accurate k-space profiles, since the k-space of motion corrupted images does not necessarily have accurate profiles, or if it does we do not know *a priori* which ones they are. To the best of our knowledge our work is the first work that has investigated the influence of motion artefacts and correction strategies on CMR segmentation.

### 3. Methods

The proposed framework of using a deep neural network for motion artefact correction on k-space data is based on a generative-adversarial network setup. Our aim is to train a successful generator to reconstruct good quality images from motion artefact corrupted k-space data similar to ([Oksuz et al., 2018a](#)) but with a novel loss function.

#### 3.1. Network Architecture

The proposed Automap-GAN algorithm follows an adversarial setup and consists of a generator and a discriminator. The generator network follows a similar architecture to ([Zhu et al., 2018](#)), which was originally developed for image reconstruction using domain specific information. In our case we additionally use a discriminator to increase the robustness and realism of the reconstructed images. The input to the network is a complex  $n$ -by- $n$  k-space matrix, which we concatenate into a  $(2 \times n \times n)$ -by-1 vector. We then use two fully connected layers: FC1 with  $2 \times n \times n$  neurons and FC2 with  $n \times n$  neurons. The output from FC2 is reshaped and two convolutional layers with 64 filters and  $5 \times 5$  filter size are used. After that a deconvolutional layer with 64 filters of size  $7 \times 7$  is applied and finally a  $1 \times 1$  layer is used to aggregate the results into an image.

The discriminator takes a generated image or a real image as input and uses two convolutional layers and a final dense layer for classification. The final output of the discriminator is a decision as to whether the generated image looks real or fake. By using outputs of the generator (artefact corrected images) and the real images from the dataset the discriminator is trained to distinguish between the artefact corrected images and high quality images. The loss function for the model uses a combination of mean squared error and structural similarity index between the predicted and real images as detailed in Section [3.2](#).

#### 3.2. Loss function

We use a combination of two loss functions following the idea proposed in ([Zhao et al., 2017](#)). The mean squared error (MSE) loss is defined as:

$$L_{\text{MSE}} = \frac{1}{N_p} \sum_{p=0}^{N_p} (I_x(p) - I_y(p))^2$$

where  $p$  denotes each pixel and  $N_p$  denotes the total number of pixels in images  $I_x$  and  $I_y$ .

Alongside this measure, we also computed the structural similarity index (SSIM) ([Wang et al., 2004](#)). SSIM has been shown to provide sensitivity to structural information and texture. The SSIM between two images is defined as follows for any image regions  $x$  and  $y$ :

$$\text{SSIM}(p) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where  $\mu_x$  and  $\mu_y$  are the average intensities for regions  $x$  and  $y$ ,  $\sigma_x$  and  $\sigma_y$  are variance values for regions  $x$  and  $y$ ,  $\sigma_{xy}$  is the covariance of regions  $x$  and  $y$  and  $c_1$  and  $c_2$  are constant values for stabilizing the denominator. The SSIM loss for every pixel  $p$  defined as:

$$L_{\text{SSIM}} = \frac{1}{N_p} \sum_{p=0}^{N_p} 1 - \text{SSIM}(p)$$

and the combined loss is defined as:

$$L_{\text{total}} = \alpha L_{\text{MSE}} + (1 - \alpha) L_{\text{SSIM}}$$

### 3.3. Implementation Details

The parameters of the convolutional and fully-connected layers were initialized randomly from a zero-mean Gaussian distribution and trained until no substantial progress was observed in the training loss. In this study, we use the RMSprop optimizer to minimize the loss. One important aspect during training is the activity regularizer, which is used after the deconvolutional layer. In our implementation, we first trained without this regularizer, finding that including it early in training led to the loss being trapped in poor local minima. Once training converged without the regularizer, it was then added, which led to the generation of sharper looking images.

The training was stopped early if no significant improvement was observed. An improvement was considered significant if the relative increase in performance was at least 0.5% over 20 epochs. To better generalize the model we applied data augmentation by rotating images in increments of 90 degrees. We also found that the success of our implementation was highly sensitive to the choice of learning rate, which we set to be  $2 \times 10^{-5}$ . The  $\alpha$  parameter balancing the MSE and SSIM losses was set at 0.6 for all experiments.

During training, a batch-size of 30 2D k-space datasets was used. We used the Keras Framework with Tensorflow backend for implementation and training the network took around 3 days on a NVIDIA Quadro 6000P GPU. Segmentation of a corrected single image sequence took around 2 seconds once the network was trained.

## 4. Experimental Results

We evaluated our algorithm on a subset of the UK Biobank dataset consisting of 148 good quality CINE CMR acquisitions. 50 temporal frames from each subject at mid-ventricular level were used to generate synthetic motion artefacts. We used 5400 2D images for training, 600 for validation and 1400 images for testing and these three sets were not over-lapping. The data were chosen to be free of other types of image quality issues such as missing axial slices and were visually verified by an expert cardiologist. The details of the acquisition protocol of the UK Biobank dataset can be found in ([Petersen et al., 2015](#)). All images were preprocessed to extract a  $128 \times 128$  pixel region and were segmented using the pre-trained CNN proposed in ([Bai et al., 2018](#)).

### 4.1. K-space corruption for synthetic data

We generated k-space corrupted data in order to simulate motion artefacts. We followed a Cartesian sampling strategy for k-space corruption to generate synthetic but realistic motion artefacts ([Oksuz et al., 2018b](#)). We first transformed each 2D short axis sequence to the Fourier domain and changed 1 in 3 Cartesian sampling lines to the corresponding lines from other cardiac phases to mimic motion artefacts. We added a random frame offset when replacing the lines. In this way the original good quality images from the training set were used to generate corresponding CMR artefact images. This is a realistic approach as the motion artefacts that occur from mis-triggering often arise from similar misplacement of k-space lines.

## 4.2. Results on synthetic dataset

**Methods of comparison:** We compared our algorithm with a variety of artefact correction strategies to produce robust segmentation results with the state-of-the-art segmentation network (Bai et al., 2018). The methods of comparison cover all possible combinations of motion artefact correction in k-space and the image domain. For image-to-image to artefact removal (i.e. post-reconstruction) we used a convolutional autoencoder (Xie et al., 2012) (CAE), a deep network based on residual learning (DNCNN) and a wide network with larger receptive fields and more channels in each layer as proposed in (Liu and Fang, 2017) (WIN5). For k-space to k-space artefact correction, we adopted the algorithm of Han and Ye (2018), which was originally proposed for accelerating MR acquisition. All of the methods used for comparison were trained with the proposed MSE/SSIM loss function. In addition to these methods we also used a variant of our algorithm, which only used MSE loss, to evaluate the influence of the combined loss function.

**Qualitative evaluation:** We show example segmentations achieved by different correction strategies together with the ground truth image in Figure 2. The improved image segmentation quality can be visualized for our proposed method in comparison to image-to-image denoising techniques.

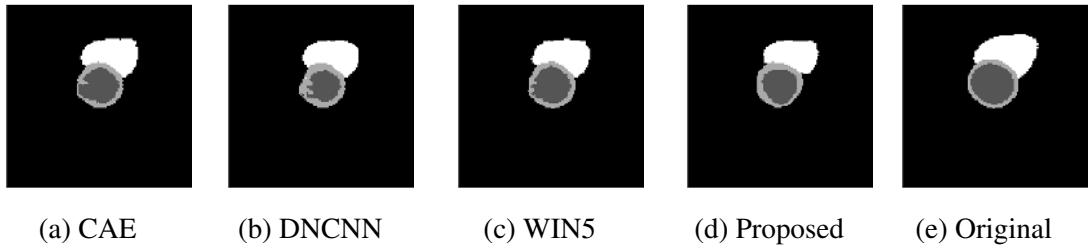


Figure 2: Synthetic dataset segmentation results after applying different artefact correction methods for an example case. CAE (a), DNCNN (b), WIN5 (c), proposed method (d) and ground truth segmentation (e). The proposed method is able to generate the best segmentation masks.

**Quantitative evaluation:** To quantitatively evaluate performance we used three metrics of left ventricular myocardium segmentation accuracy. We first computed the Dice overlap measure, which is defined between two regions  $A$  and  $B$  as:

$$D(A, B) = \frac{2\|A \cap B\|}{\|A\| \cup \|B\|}.$$

We also computed the mean contour distance (MCD) and Hausdorff distance (HD) to evaluate the mean and the maximum distances respectively between the segmentation contours  $C_A$  and  $C_B$ . These are defined as,

$$MCD(A, B) = \frac{1}{2C_A} \sum_{p \in C_A} d(p, C_B) + \frac{1}{2C_B} \sum_{p \in C_B} d(p, C_A)$$

$$HD(C_A, C_B) = \max \left\{ \max_{a \in C_A} \min_{b \in C_B} d(a, b), \min_{b \in C_B} \max_{a \in C_A} d(a, b) \right\}$$

Table 1: Mean Dice, MCD and HD results of segmentation after motion artefact correction with various methods using the network output as ground truth.

	Dice	MCD	HD
Corrupted	$0.635 \pm 0.027$	$2.583 \pm 1.078$	$6.213 \pm 2.712$
K-space ( <a href="#">Han and Ye, 2018</a> )	$0.645 \pm 0.038$	$2.214 \pm 1.205$	$5.971 \pm 2.587$
CAE ( <a href="#">Xie et al., 2012</a> )	$0.653 \pm 0.041$	$2.109 \pm 1.011$	$5.602 \pm 2.210$
DNCNN ( <a href="#">Zhang et al., 2017</a> )	$0.662 \pm 0.034$	$2.071 \pm 0.832$	$5.582 \pm 2.001$
WIN5 ( <a href="#">Liu and Fang, 2017</a> )	$0.681 \pm 0.018$	$1.966 \pm 0.818$	$5.404 \pm 1.818$
Proposed-MSE ( <a href="#">Oksuz et al., 2018a</a> )	$0.679 \pm 0.015$	$1.932 \pm 0.918$	$4.968 \pm 1.718$
Proposed-Combined Loss	<b><math>0.722 \pm 0.014</math></b>	<b><math>1.829 \pm 0.914</math></b>	<b><math>4.811 \pm 1.521</math></b>

where  $d$  represents the distance between points  $a \in C_A$  and  $b \in C_B$ . Note that we report results for the myocardial region only for brevity. Similar results were observed for the right and left ventricle cavities.

Table 1 shows the Dice segmentation accuracies for the segmentation masks produced by the segmentation network using the output of each image artefact correction algorithm. For these experiments the ground truth was the myocardium segmentation that was generated using the network on the original uncorrupted data. The proposed adversarial Automap technique is capable of correcting motion artefacts and enabling high segmentation accuracy compared to the other techniques in both mid-ventricular slices and over all slices. We can also see that the image-to-image denoising techniques have enabled better segmentation performance compared to k-space based artefact correction. The combined MSE and SSIM loss improves segmentation performance compared to the MSE loss alone.

In Table 2, we report the Dice, MCD and HD scores, but this time using as ground truth the expert annotated masks at end-systole and end-diastole phases. Again, the adversarial Automap technique was capable of correcting motion artefacts and enabling high accuracy of segmentation compared to the other techniques. In general, segmentation performance was a bit lower compared to Table 1. This is due to the fact that the ground truth segmentations have a mean Dice score of 0.912 with the segmentations from the original images (i.e. produced by ([Bai et al., 2018](#))) as illustrated in Table 2.

#### 4.3. Qualitative results on real motion artefact case

To illustrate the performance of our technique on artefact correction, we applied it to a dataset from the UK Biobank containing real mis-triggering artefacts (i.e. not synthetically corrupted). This dataset could not be segmented by the expert cardiologists due to severe motion artefacts. The visual segmentation and image quality results are illustrated in Figure 3, which shows improved image quality and CMR segmentation with the network especially in the left ventricular blood pool.

### 5. Discussion and Conclusion

In this paper, we have evaluated the influence of a variety of image artefact correction mechanisms on CMR segmentation accuracy. We have proposed an architecture with a combined MSE and

Table 2: Mean Dice, MCD and HD results of segmentation after motion artefact correction with various methods using the expert annotated mask as ground truth.

	Dice	MCD	HD
Original Images (Bai et al., 2018)	$0.912 \pm 0.027$	$0.876 \pm 0.497$	$2.583 \pm 0.625$
Corrupted	$0.590 \pm 0.053$	$2.691 \pm 1.184$	$6.587 \pm 2.018$
K-space (Han and Ye, 2018)	$0.613 \pm 0.032$	$2.304 \pm 1.004$	$6.021 \pm 1.907$
CAE (Xie et al., 2012)	$0.641 \pm 0.040$	$2.184 \pm 1.118$	$5.819 \pm 1.820$
DNCNN (Zhang et al., 2017)	$0.657 \pm 0.047$	$2.182 \pm 0.972$	$5.712 \pm 1.576$
WIN5 (Liu and Fang, 2017)	$0.676 \pm 0.038$	$2.032 \pm 0.808$	$5.481 \pm 1.471$
Proposed-MSE (Oksuz et al., 2018a)	$0.671 \pm 0.037$	$2.018 \pm 0.701$	$5.261 \pm 1.487$
Proposed-Combined Loss	<b><math>0.696 \pm 0.027</math></b>	<b><math>1.958 \pm 0.674</math></b>	<b><math>5.161 \pm 1.107</math></b>

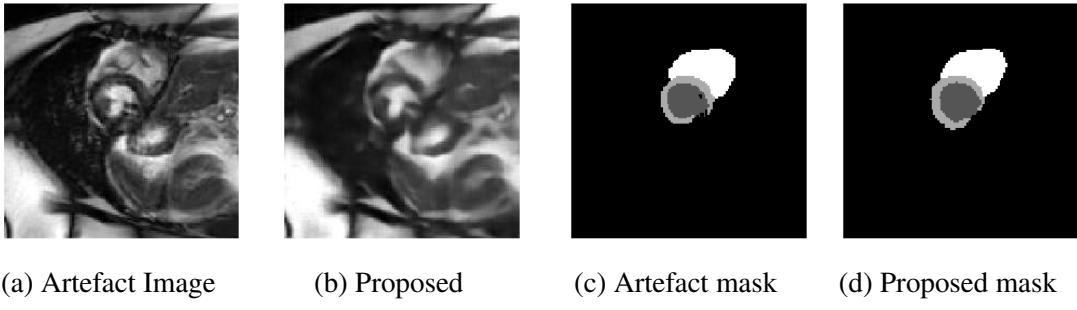


Figure 3: Example of a mis-triggering artefact and its segmentation from the UK Biobank dataset. Artefact image (a), corrected image with proposed method (b), segmentation of artefact image (c) and segmentation after proposed correction (d). The proposed method is able to correct the motion artefacts and improve segmentation especially inside the left ventricular blood pool. (The ground truth segmentation does not exist for this dataset.)

SSIM loss to achieve high segmentation accuracy. We have shown that low image quality has a deteriorating influence on image segmentation pipelines, which should be taken into account. Our method has great potential to address the problem of low segmentation accuracy on clinical data, which is likely to contain more motion corruption compared to the UK Biobank data. One limitation is the high memory requirements of our technique to correct motion artefacts.

Future work will involve end-to-end image artefact correction and segmentation mechanisms to improve the image quality and segmentation accuracy simultaneously. Moreover, we will investigate alternative implementations of our algorithm to reduce the memory requirements.

In conclusion, our work opens up the path for a reconsideration of common assumptions about how well segmentation accuracy transfers to clinical data ‘in the wild’.

## Acknowledgments

This work was supported by an EPSRC programme Grant (EP/P001009/1) and the Wellcome EP-SRC Centre for Medical Engineering at the School of Biomedical Engineering and Imaging Sci-

ences, King's College London (WT 203148/Z/16/Z). This research has been conducted using the UK Biobank Resource under Application Number 17806. The GPU used in this research was generously donated by the NVIDIA Corporation.

## References

- Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M. Lee, Nay Aung, Elena Lukaschuk, Mihir M. Sanghvi, Filip Zemrak, Kenneth Fung, Jose Miguel Paiva, Valentina Carapella, Young Jin Kim, Hideaki Suzuki, Bernhard Kainz, Paul M. Matthews, Steffen E. Petersen, Stefan K. Piechnik, Stefan Neubauer, Ben Glocker, and Daniel Rueckert. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20(1):65, Sep 2018.
- Pedro F Ferreira, Peter D Gatehouse, Raad H Mohiaddin, and David N Firmin. Cardiovascular magnetic resonance artefacts. *Journal of Cardiovascular Magnetic Resonance*, 15(1):41, 2013.
- Yoseob Han and Jong Chul Ye. k-space deep learning for accelerated mri. *arXiv preprint arXiv:1805.03779*, 2018.
- Andreas Hauptmann, Simon Arridge, Felix Lucka, Vivek Muthurangu, and Jennifer A Steeden. Real-time cardiovascular mr with spatio-temporal artifact suppression using deep learning—proof of concept in congenital heart disease. *Magnetic resonance in medicine*, 81(2):1143–1156, 2019.
- Yoon-Chul Kim, Jon-Fredrik Nielsen, and Krishna S Nayak. Automatic correction of echo-planar imaging (epi) ghosting artifacts in real-time interactive cardiac mri using sensitivity encoding. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(1):239–245, 2008.
- Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60 – 88, 2017.
- Peng Liu and Ruogu Fang. Wide inference network for image denoising via learning pixel-distribution prior. *arXiv preprint arXiv:1707.05414*, 2017.
- Jyrki Lötjönen, Mika Pollari, Sari Kivistö, and Kirsi Lauerma. Correction of motion artifacts from cardiac cine magnetic resonance images1. *Academic radiology*, 12(10):1273–1284, 2005.
- Ilkay Oksuz, James Clough, Aurelien Bustin, Gastao Cruz, Claudia Prieto, Rene Botnar, Daniel Rueckert, Julia A. Schnabel, and Andrew P. King. Cardiac mr motion artefact correction from k-space using deep learning-based reconstruction. In Florian Knoll, Andreas Maier, and Daniel Rueckert, editors, *Machine Learning for Medical Image Reconstruction*, pages 21–29, Cham, 2018a. Springer International Publishing. ISBN 978-3-030-00129-2.
- Ilkay Oksuz, Bram Ruijsink, Esther Puyol-Antón, Aurelien Bustin, Gastao Cruz, Claudia Prieto, Daniel Rueckert, Julia A. Schnabel, and Andrew P. King. Deep learning using k-space based data augmentation for automated cardiac mr motion artefact detection. In Alejandro F. Frangi,

Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 250–258, Cham, 2018b. Springer International Publishing. ISBN 978-3-030-00928-1.

Steffen E Petersen, Paul M Matthews, Jane M Francis, Matthew D Robson, Filip Zemrak, Redha Boubertakh, Alistair A Young, Sarah Hudson, Peter Weale, Steve Garratt, et al. Uk biobank’s cardiovascular magnetic resonance protocol. *Journal of cardiovascular magnetic resonance*, 18(1):8, 2015.

Farhood Saremi, John D Grizzard, and Raymond J Kim. Optimizing cardiac mr imaging: practical remedies for artifacts. *Radiographics*, 28(4):1161–1187, 2008.

Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503, 2018a.

Jo Schlemper, Ozan Oktay, Wenjia Bai, Daniel C Castro, Jinming Duan, Chen Qin, Jo V Hajnal, and Daniel Rueckert. Cardiac mr segmentation from undersampled k-space using deep latent representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 259–267. Springer, 2018b.

Muhan Shao, Shuo Han, Aaron Carass, Xiang Li, Ari M Blitz, Jerry L Prince, and Lotta M Ellingsen. Shortcomings of ventricle segmentation using deep convolutional networks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 79–86. Springer, 2018.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems*, pages 341–349, 2012.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017.

Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487, 2018.

# Weakly Supervised Deep Nuclei Segmentation using Points Annotation in Histopathology Images

**Hui Qu**<sup>1</sup>

HUI.QU@CS.RUTGERS.EDU

**Pengxiang Wu**<sup>1</sup>

PW241@CS.RUTGERS.EDU

**Qiaoying Huang**<sup>1</sup>

QH55@CS.RUTGERS.EDU

**Jingru Yi**<sup>1</sup>

JY486@CS.RUTGERS.EDU

**Gregory M. Riedlinger**<sup>2</sup>

GR338@CINJ.RUTGERS.EDU

**Subhajyoti De**<sup>2</sup>

SD948@CINJ.RUTGERS.EDU

**Dimitris N. Metaxas**<sup>1</sup>

DNM@CS.RUTGERS.EDU

<sup>1</sup> Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA.

<sup>2</sup> Rutgers Cancer Institute, 195 Little Albany St, New Brunswick, NJ 08901, USA.

## Abstract

Nuclei segmentation is a fundamental task in histopathological image analysis. Typically, such segmentation tasks require significant effort to manually generate pixel-wise annotations for fully supervised training. To alleviate the manual effort, in this paper we propose a novel approach using points only annotation. Two types of coarse labels with complementary information are derived from the points annotation, and are then utilized to train a deep neural network. The fully-connected conditional random field loss is utilized to further refine the model without introducing extra computational complexity during inference. Experimental results on two nuclei segmentation datasets reveal that the proposed method is able to achieve competitive performance compared to the fully supervised counterpart and the state-of-the-art methods while requiring significantly less annotation effort. Our code is publicly available<sup>1</sup>.

**Keywords:** Nuclei segmentation, Weak supervision, Deep learning, Voronoi diagram, Conditional random field.

## 1. Introduction

Nuclei segmentation is a critical step in the automatic analyses of histopathology images, because the nuclear features such as average size, density and nucleus-to-cytoplasm ratio are often related to the clinical diagnosis and management of cancer. Modern deep learning based nuclei segmentation methods (Xing et al., 2016; Kumar et al., 2017; Naylor et al., 2017, 2018; Mahmood et al., 2018; Janowczyk and Madabhushi, 2016; Qu et al., 2019) have achieved better performance than traditional approaches such as watershed segmentation (Veta et al., 2013) and graph-based segmentation (Al-Kofahi et al., 2010). However, the fully supervised training of deep neural networks in these methods requires a large amount of pixel-wise annotated data, which are difficult to collect because assigning a nucleus/background class label to every pixel in the image is time-consuming and requires specific domain knowledge. Therefore, methods using weak annotations are needed to reduce the annotation burden.

---

1. The code can be found at: <https://github.com/huiqu18/WeaklySegPointAnno>

There have been various methods using weak annotations in image segmentation. For natural images, weak annotations include image-level tags (Papandreou et al., 2015; Pathak et al., 2015), scribbles (Lin et al., 2016), points (Bearman et al., 2016) and bounding boxes (Dai et al., 2015; Khoreva et al., 2017; Rajchl et al., 2017). Image-level tags are the class information of objects, which are not used in medical image segmentation where object classes in images are usually fixed (e.g., nuclei and background in our task). Scribbles annotation, which requires at least one scribble for every object, is not suitable for our task due to the small size and large number of nuclei. The objectiveness prior in the points supervision work (Bearman et al., 2016) is not working here since nuclei are small and thus the prior is inaccurate. Bounding boxes are more well defined and are also commonly adopted in medical images (Yang et al., 2018; Zhao et al., 2018). However, it is still time-consuming and difficult to label an image using bounding boxes for hundreds of nuclei, especially when the density is high. Kervadec et al. (Kervadec et al., 2019) used a small fraction of full labels and imposed a size constraint in the loss function, which achieved good performance but is not applicable for multiple objects of a same class. Different from existing methods, in this work we propose to employ points annotation for nuclei segmentation. All a pathologist needs to do is mark the location of every nucleus with a point. Our method is efficient and more annotation-friendly, and to the best of our knowledge, this is the first time points annotation has been successfully applied to nuclei segmentation.

In practice, the points annotation itself is not sufficient to directly supervise the training of neural networks. To address this problem, we take advantage of the original image and the shape prior of nuclei to derive two types of coarse labels from the points annotation using the Voronoi diagram and the  $k$ -means clustering algorithm. The Voronoi diagram was ever used in nuclei detection (Kost et al., 2017) for training sample selection, but here we utilize it to generate the coarse labels for nuclei segmentation, which is a different and much harder task. These two types of coarse labels are then used to train a deep convolutional neural network (CNN) with the cross entropy loss.

A common problem in various weakly supervised segmentation tasks is that the key information near the object boundaries is missing. Therefore, post-processing like the dense conditional random field (CRF) (Chen et al., 2015) or graph search (Yang et al., 2018) is needed to refine the object boundaries, at the expense of increased processing time. Inspired by Tang et al.’s work (Tang et al., 2018), we utilize the dense CRF in the loss function to fine-tune the trained model rather than add a post-processing step, thereby leading to a more efficient model as the loss is no longer needed during inference. This property makes our method more preferable in nuclei segmentation of large Whole Slide Images.

In summary, the contributions of our work include:

- To the best of our knowledge, we are the first to successfully utilize the points annotation for nuclei segmentation in histopathology images.
- We present a new method for deriving two types of informative pixel-level labels from points label using the Voronoi diagram and  $k$ -means clustering algorithm, and employ the dense CRF loss for model refinement in nuclei segmentation.
- We show that our approach achieves competitive segmentation performance on two nuclei segmentation datasets. The accuracy is comparable to that obtained with full supervised approaches.

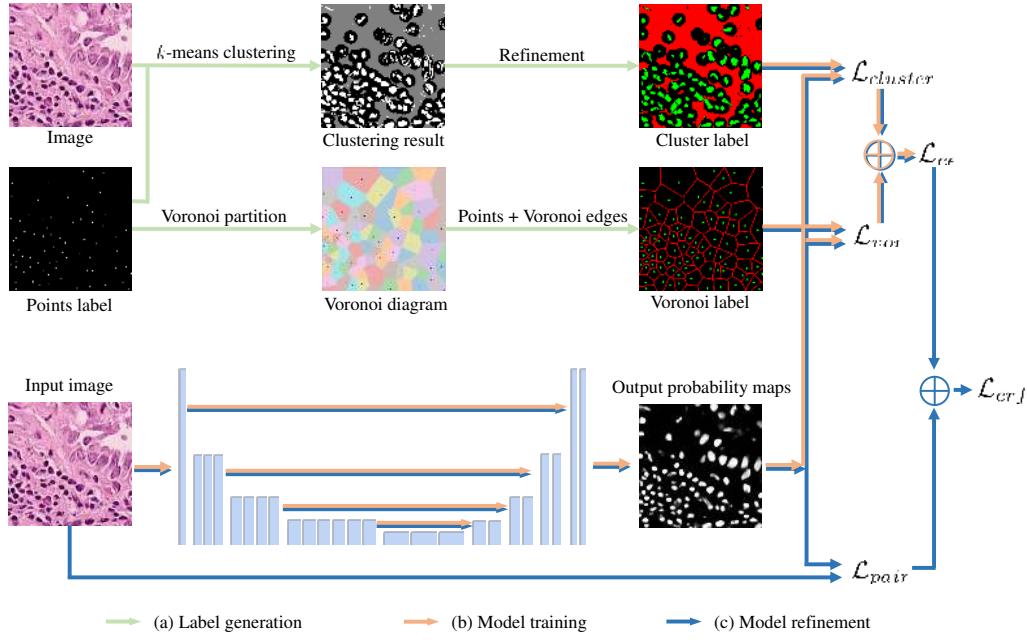


Figure 1: Overview of the proposed approach. (a) Label generation. The Voronoi label and cluster label are generated using the points label and original image. The green, red and black colors indicate nuclei, background and ignored pixels, respectively. (b) Model training using the cross entropy loss. (c) Model refinement using the CRF loss.

## 2. Methods

In this section we describe our approach in detail. In particular, our point-level supervision for training a nuclei segmentation model consists of three parts: (1) coarse pixel-level labels generation using points annotation; (2) segmentation network training with coarse labels; (3) model refinement using the dense CRF loss.

### 2.1. From point-level to pixel-level labels

The point-level labels cannot be used directly for the training of a CNN with the cross entropy loss due to the lack of (negative) background labels since all annotated points belong to the (positive) nuclei category. To solve this issue, the first step is to exploit the information we have to generate useful pixel-level labels for both classes. We have the following observations: (1) Each point is expected to be located or close to the center of a nucleus, and the shapes of most nuclei are nearly ellipses, i.e., they are convex. (2) The colors of nuclei pixels are often different from the surrounding background pixels. Based on these observations, we propose to utilize the Voronoi diagram and  $k$ -means clustering methods to produce two types of pixel-level labels.

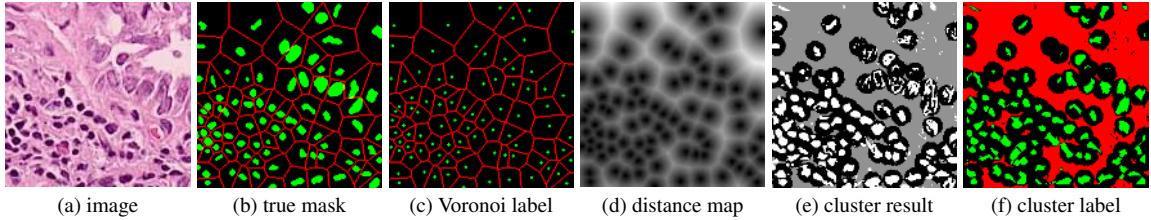


Figure 2: Label generation. (a) original image, (b) ground-truth nuclei masks (in green) and Voronoi edges (in red), (c) Voronoi label, (d) distance map, (e) clustering result, (f) cluster label (green: nuclei, red: background, black: ignored).

### 2.1.1. VORONOI LABELS

Voronoi diagram is a partitioning of a plane into convex polygons (Voronoi cells) according to the distance to a set of points in the plane. There is exactly one point (seed point) in each cell and all points in a cell are closer to its seed point than other seed points. In our task, the annotated points of an image can be treated as seed points to calculate the Voronoi diagram, see Fig. 1. For each cell, assuming that the corresponding nucleus is located within the cell, then the Voronoi edges separate all nuclei well and the edge pixels belong to the background. This assumption holds for most of the nuclei because the points are around the centers and nuclear shapes are nearly convex (Fig. 2(b)).

Treating the Voronoi edges as background pixels and the annotated points (dilated with a disk kernel of radius 2) as nuclei pixels, we obtain the Voronoi point-edge label (Fig. 2(c)). All other pixels are ignored during training. Note that although the pixels on the Voronoi edge between two touching nuclei may not necessarily be background, the edges are still helpful in guiding the network to separate the nuclei. The Voronoi labels aim to segment the central parts of nuclei and are not able to extract the full masks, because they lack the information of nuclear boundaries and shapes. To overcome the weakness, we generate another kind of labels that contain this information as a complement.

### 2.1.2. CLUSTER LABELS

Considering the difference in colors between nuclei and background pixels, it is feasible to perform a rough segmentation using clustering methods. We choose the  $k$ -means clustering algorithm to extract both nuclei and background pixels from the original image, and produce the cluster labels based on the results. Given an image  $\mathbf{x}$  with  $N$  pixels  $(x_1, x_2, \dots, x_N)$ ,  $k$ -means clustering aims to partition the  $N$  pixels into  $k$  clusters  $\mathcal{S} = (S_1, S_2, \dots, S_k)$  according to the feature vector  $f_{x_i}$  of each pixel  $x_i$ , such that the sum of within-cluster variances is minimized :

$$\arg \min_{\mathcal{S}} \sum_{i=1}^k \sum_{x \in S_i} \|f_x - c_i\|^2. \quad (1)$$

We use  $k$ -means to divide all pixels into  $k = 3$  clusters: nuclei, background and ignored. The cluster that has maximum overlap with points label is considered as nuclei, and the cluster that has minimum overlap with the dilated points label is considered as background. The remaining one is the ignored class. The pixels of ignored class are often located around the nuclear boundaries, which are hard for a clustering method to assign correct labels.

For the feature vector  $f$ , color is the straightforward choice. However, clustering with color will result in wrong assignments for pixels inside some nuclei that have non-uniform colors. To cope with this issue, we propose to add a distance value in the feature vector. In a distance map (Fig. 2(d)), each value indicates the distance of that pixel to the closest nuclear point and therefore incorporates the position information. In particular, the pixels that belong to nuclei should be close enough to points in the label while background pixels are expected to be relatively far from those points. The distance map can be calculated by the distance transform of the complement image of points label. Combining the distance value  $d_i$  with the RGB color values  $(r_i, g_i, b_i)$  as the feature vector  $f_{x_i} = (\hat{d}_i, \hat{r}_i, \hat{g}_i, \hat{b}_i)$  and performing  $k$ -means clustering, we obtain the initial cluster labels (Fig. 2(e)).  $\hat{d}_i$  is the clipped value by truncating large values to 20 and  $\hat{r}_i, \hat{g}_i, \hat{b}_i$  are scaled color values such that each element in the feature vector has similar range. The final cluster label (Fig. 2(f)) is generated by refining the clustering result with morphological opening operation. The cluster labels have more shape information about the nuclei compared to Voronoi label, but may contain more errors and uncertainties. We argue that these two types of labels are complementary to each other and would jointly lead to better results.

## 2.2. Training deep neural networks with pixel-level labels

Once we have the pixel-level labels, we are able to train a deep convolutional neural network for nuclei segmentation. The network (shown in Fig. 1) we use is a modified version of U-net (Ronneberger et al., 2015). We replace the encoder part of U-net with the convolution layers of ResNet34 (He et al., 2016), which is more powerful in representation ability and can be initialized with pretrained parameters from image classification task on ImageNet (Russakovsky et al., 2015). The network outputs two probability maps of background and nuclei, which are used to calculate two cross entropy losses with respect to the cluster label  $\mathcal{L}_{cluster}$  and Voronoi label  $\mathcal{L}_{vor}$ :

$$\mathcal{L}_{cluster/vor}(\mathbf{y}, \mathbf{t}) = -\frac{1}{|\Omega|} \sum_{i \in \Omega} [t_i \log y_i + (1 - t_i) \log(1 - y_i)], \quad (2)$$

where  $\mathbf{y}$  is the probability map,  $\mathbf{t}$  is the cluster label or Voronoi label, and  $\Omega$  is the set consisting of non-ignored pixels. The final loss is  $\mathcal{L}_{ce} = \mathcal{L}_{cluster} + \mathcal{L}_{vor}$ .

## 2.3. Model refinement using dense CRF loss

The model trained using the two types of labels is able to predict the masks of individual nuclei with high accuracy. To further improve the performance, we refine the nuclear boundaries with the dense CRF loss. Previously post-processing such as region growing (Kumar et al., 2017), graph search (Yang et al., 2018) or dense CRF (Chen et al., 2015) is often utilized to refine the segmentation results. These algorithms introduce more computational complexity, making them unsuitable for the processing of large resolution Whole Slide Images. To solve this problem, similar to (Tang et al., 2018) we embed the dense CRF into the loss function to improve the accuracy. The loss function is not calculated during inference, and therefore will not introduce additional computational cost after training.

Let  $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_N)$  denote the predicted label (0 for background and 1 for nuclei) from probability maps  $\mathbf{y}$  and  $\mathbf{t}$  be the label. The dense CRF is to minimize the energy function:

$$E(\tilde{\mathbf{y}}, \mathbf{t}) = \sum_i \phi(\tilde{y}_i, t_i) + \sum_{i,j} \psi(\tilde{y}_i, \tilde{y}_j), \quad (3)$$

where  $\phi$  is the unary potential that measures how likely a pixel belongs to a certain class, and  $\psi$  is the pairwise potential that measures how different a pixel's label is from all other pixels' in the image. The unary term is replaced with the cross entropy loss  $\mathcal{L}_{ce}$ . The pairwise potential usually has the form:

$$\psi(\tilde{y}_i, \tilde{y}_j) = \mu(\tilde{y}_i, \tilde{y}_j) W_{ij} = \mu(\tilde{y}_i, \tilde{y}_j) \sum_{m=1}^K w_m k_m(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_j), \quad (4)$$

where  $\mu$  is a label compatibility function,  $W_{ij}$  is the affinity between pixels  $i, j$  and is often calculated by the sum of Gaussian kernels  $k_m$ . In this work we choose  $\mu$  as the Potts model, i.e.,  $\mu(\tilde{y}_i, \tilde{y}_j) = [\tilde{y}_i \neq \tilde{y}_j]$ , and bilateral feature vector  $\tilde{\mathbf{f}}_i = \left( \frac{p_i}{\sigma_{pq}}, \frac{q_i}{\sigma_{pq}}, \frac{r_i}{\sigma_{rgb}}, \frac{g_i}{\sigma_{rgb}}, \frac{b_i}{\sigma_{rgb}} \right)$  that contains both location and color information.  $\sigma_{pq}$  and  $\sigma_{rgb}$  are Gaussian bandwidth.

To adapt the energy function to a loss function that is differentiable for training, we relax the pairwise potential as (Tang et al., 2018):  $\psi(\tilde{y}_i, \tilde{y}_j) = \tilde{y}_i(1 - \tilde{y}_j)W_{ij}$ . Therefore, the dense CRF loss can be expressed as:

$$\mathcal{L}_{crf}(\mathbf{y}, \mathbf{t}_{cluster}, \mathbf{t}_{vor}) = \mathcal{L}_{ce}(\mathbf{y}, \mathbf{t}_{cluster}, \mathbf{t}_{vor}) + \beta \mathcal{L}_{pair}(\mathbf{y}), \quad (5)$$

where  $\mathcal{L}_{pair}(\mathbf{y}) = \sum_{i,j} y_i(1 - y_j)W_{ij}$  is the pairwise potential loss and  $\beta$  is the weighting factor. The CRF loss is used to fine-tune the trained model. Due to the large number of pixels in an image, the cost of directly computing the affinity matrix  $W = [W_{ij}]$  is prohibitive. For instance, there are  $N^2 = 1.6 \times 10^9$  elements in  $W$  for an image of size  $200 \times 200$  that has  $N = 40000$  pixels. We adopt fast mean-field inference based on high-dimensional filtering (Adams et al., 2010) to compute the pairwise potential part.

### 3. Experiments and Results

To validate our method, we apply it to two datasets of H&E stained histopathology images for nuclei segmentation and compare the results with fully supervised methods, including the same model trained with full masks, the CNN3 method proposed by Kumar et al. (Kumar et al., 2017) and the DIST method proposed by Naylor et al. (Naylor et al., 2018).

#### 3.1. Datasets, evaluation and implementation details

**Datasets** The Lung Cancer dataset contains 40 images from 8 different lung cancer cases, and each case has 5 images of size about  $900 \times 900$ . These images are split into train, validation and test sets, consisting of 24, 8 and 8 images, respectively. Each set has at least one image of each case. Another dataset is publicly available, i.e., MultiOrgan dataset (Kumar et al., 2017). It consists of 30 image of size  $1000 \times 1000$ , which are taken from multiple hospitals and include a diversity of nuclear appearances from seven organs (Kumar et al., 2017). Both datasets have full mask annotation. We obtain the points annotation for the training sets by computing the central point of each nuclear mask.

**Evaluation metrics** Four metrics are used for evaluation, including pixel accuracy, pixel-level F1 score, object-level Dice coefficient (Sirinukunwattana et al., 2015) and the Aggregated Jaccard Index (AJI) (Kumar et al., 2017). The pixel-level F1 score is defined as  $F1 = 2 \cdot TP / (2 \cdot TP + FP + FN)$ , where TP, FP, FN are the numbers of true positive, false positive and false negative pixels,

Table 1: Results on Lung Cancer dataset using our methods in different settings.

Method	Pixel-level		Object-level	
	Acc	F1	Dice <sub>obj</sub>	AJI
Full	0.9615	0.8771	0.8521	0.6979
Weak/Voronoi	0.9147	0.6596	0.6472	0.4791
Weak/Cluster	0.9188	0.7662	0.5936	0.2332
Weak w/o CRF	0.9413	0.8028	0.7885	0.6328
Weak w/ CRF	<b>0.9433</b>	<b>0.8120</b>	<b>0.8002</b>	<b>0.6503</b>

respectively. The object-level Dice coefficient is defined as

$$Dice_{obj}(\mathcal{G}, \mathcal{S}) = \frac{1}{2} \left[ \sum_{i=1}^{n_g} \gamma_i Dice(G_i, S^*(G_i)) + \sum_{j=1}^{n_s} \sigma_j Dice(G^*(S_j), S_j) \right] \quad (6)$$

where  $\gamma_i$ ,  $\sigma_j$  are the weights related to object areas,  $\mathcal{G}$ ,  $\mathcal{S}$  are the set of ground-truth objects and segmented objects,  $S^*(G_i)$ ,  $G^*(S_j)$  are the segmented object that has maximum overlapping area with  $G_i$  and ground-truth object that has maximum overlapping area with  $S_j$ , respectively. The correspondence is built if the overlap area of two objects are more than 50%. This metric takes into account each object individually, and measures how well each segmented object overlaps with the ground truth objects, as well as how well each ground truth object overlaps the segmented objects (Sirinukunwattana et al., 2015). Another object-level metric AJI is proposed to evaluate the performance in nuclei segmentation and defined as

$$AJI = \frac{\sum_{i=1}^{n_g} |G_i \cap S(G_i)|}{\sum_{i=1}^{n_g} |G_i \cup S(G_i)| + \sum_{k \in K} |S_k|} \quad (7)$$

where  $S(G_i)$  is the segmented object that has maximum overlap with  $G_i$  with regard to Jaccard index,  $K$  is the set containing segmentation objects that have not been assigned to any ground-truth object.

**Implementation details** Color normalization (Reinhard et al., 2001) is applied to all images to remove color variations caused by staining. Due to the small size of datasets, data augmentation such as random crop, scale, rotation, flipping, and affine transformation are adopted. The network is initialized with pretrained parameters and updated using the Adam optimizer. In weakly supervised settings, we train a model for 60 epochs with a learning rate of 1e-4, and fine-tune the model using dense CRF loss for 10 epochs with a learning rate of 1e-5. The parameters in CRF loss are  $\sigma_{pq} = 10$ ,  $\sigma_{rgb} = 10$ ,  $\beta = 0.0005$ . The validation set is not used because we have no access to ground-truth masks when training with points label. In fully supervised settings, we train 200 epochs using binary masks with a learning rate of 1e-4. The validation set is used to select the best model for test.

### 3.2. Results and comparison

**The effects of two types of labels** In order to show the importance of two types of generated labels, we report the results using either type of labels on the Lung Cancer dataset in Table 1. Compared to the results using the cluster labels, those with Voronoi labels are better in the object-level

Table 2: Results on MultiOrgan dataset for CNN3 (Kumar et al., 2017), DIST (Naylor et al., 2017), fully supervised training and our methods with and without CRF loss.

Method	Pixel-level		Object-level	
	Acc	F1	Dice <sub>obj</sub>	AJI
CNN3	-	-	-	0.5083
DIST	-	0.7623	-	<b>0.5598</b>
Full	0.9194	0.8100	0.6763	0.3919
Weak w/o CRF	0.9052	0.7745	0.7231	0.5045
Weak w/ CRF	<b>0.9071</b>	<b>0.7776</b>	<b>0.7270</b>	0.5097

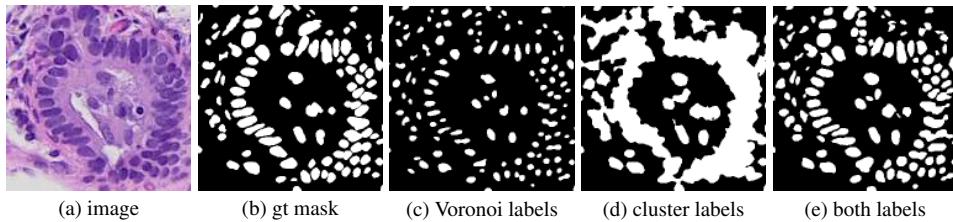


Figure 3: Results using different pixel-level labels: (a) image, (b) ground-truth mask, (c)-(e) are results using Voronoi labels, cluster labels and both labels, respectively.

metrics but worse in pixel-level metrics. This is because the model trained with Voronoi labels predicts the central parts of nuclei, resulting in small separated instances (Fig. 3(c)). While lacking the Voronoi edge information, the model using cluster labels is not able to separate close nuclei (Fig. 3(d)). In contrast, segmentation results using both labels are better than those with either label alone (Fig. 3(e))).

**The effects of dense CRF loss** From Table 1, it can be observed that the refinement with dense CRF loss improves the segmentation performance on the Lung Cancer dataset for all four metrics, but it is less effective on the MultiOrgan dataset. The reason is that in the MultiOrgan dataset there are many more crowded and touching nuclei that have no clear boundaries. CRF loss cannot handle these hard cases well.

**Comparison to fully supervised methods** The segmentation performance of our weakly supervised method is close to that of the fully supervised models with the same network structure. On the Lung Cancer dataset, the gaps for accuracy, F1 score, Dice and AJI are 1.9%, 7.4%, 6.1%, 6.8%, respectively. On the MultiOrgan dataset, the gaps for accuracy and F1 score are 1.3% and 4.0%. However, the fully supervised model has very low Dice and AJI, since for fair comparison we didn't perform post-processing to separate the touching nuclei for any of the methods. The weakly supervised model is able to separate most of them due to the Voronoi labels while the fully supervised model failed to achieve this. Compared to the CNN3 method in (Kumar et al., 2017), our method achieved the similar accuracy in terms of the AJI value. Compared to the state-of-the-art DIST method (Naylor et al., 2018), our approach has the higher pixel-level F1 score, but still has room

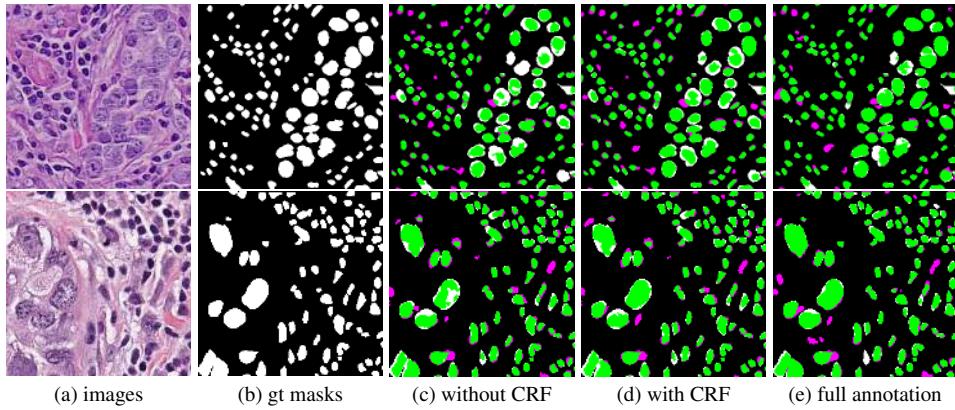


Figure 4: Comparison of weakly and fully supervised training: (a) images, (b) ground-truth masks, (c)-(e) are results for weak labels without, with CRF loss and full labels, respectively, overlapped with ground-truth masks. Pixels in green, magenta, white are true positives, false positives and false negatives, respectively.

for improvement on the nuclear shapes, as indicated by the AJI values. Several image results are illustrated in Fig. 4.

**Annotation time** In order to show the time efficiency of points annotation, our pathologist annotated eight images (one per case) in the Lung Cancer dataset using points, bounding boxes and full masks, respectively. The average time spent on each image (about 600 nuclei in average) for full masks is 115 minutes while for bounding boxes, 67 minutes. However, it only takes about 14 minutes for points annotation.

#### 4. Conclusion

In this paper we present a new weakly supervised nuclei segmentation method using only points annotation. We generate the Voronoi label and cluster label from the points label and take advantage of the dense CRF loss to refine our trained model. Our method is able to achieve comparable performance as fully supervised methods while requiring much less annotation effort which in turn allows us to analyze large amounts of data.

#### References

- Andrew Adams, Jongmin Baek, and Myers Abraham Davis. Fast high-dimensional filtering using the permutohedral lattice. *Computer Graphics Forum*, 29(2):753–762, 2010.
- Yousef Al-Kofahi, Wiem Lassoued, William Lee, and Badrinath Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering*, 57(4):841–852, 2010.
- Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What’s the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision*, pages 549–565. Springer, 2016.

- Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643, 2015.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7, 2016.
- Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. Constrained-cnn losses for weakly supervised segmentation. *Medical image analysis*, 54:88–99, 2019.
- Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, page 3, 2017.
- Henning Kost, André Homeyer, Jesper Molin, Claes Lundström, and Horst Karl Hahn. Training nuclei detection algorithms with simple annotations. *Journal of Pathology Informatics*, 8, 2017.
- Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7):1550–1560, 2017.
- Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016.
- Faisal Mahmood, Daniel Borders, Richard Chen, Gregory N McKay, Kevan J Salimian, Alexander Baras, and Nicholas J Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *arXiv preprint arXiv:1810.00236*, 2018.
- Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Nuclei segmentation in histopathology images using deep neural networks. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 933–936. IEEE, 2017.
- Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging*, 2018.
- George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015.

- Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015.
- Hui Qu, Gregory Riedlinger, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Subhajyoti De, and Dimitris Metaxas. Joint segmentation and fine-grained classification of nuclei in histopathology images. In *International Symposium on Biomedical Imaging*, pages 900–904. IEEE, 2019.
- Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A Rutherford, Joseph V Hajnal, Bernhard Kainz, et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE transactions on medical imaging*, 36(2):674–683, 2017.
- Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- Korsuk Sirinukunwattana, David RJ Snead, and Nasir M Rajpoot. A stochastic polygons model for glandular structures in colon histology images. *IEEE transactions on Medical Imaging*, 34(11):2366–2378, 2015.
- Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018.
- Mitko Veta, Paul J Van Diest, Robert Kornegoor, André Huisman, Max A Viergever, and Josien PW Pluim. Automatic nuclei segmentation in h&e stained breast cancer histopathology images. *PloS one*, 8(7):e70221, 2013.
- Fuyong Xing, Yuanpu Xie, and Lin Yang. An automatic learning-based framework for robust nucleus segmentation. *IEEE transactions on Medical Imaging*, 35(2):550–566, 2016.
- Lin Yang, Yizhe Zhang, Zhuo Zhao, Hao Zheng, Peixian Liang, Michael TC Ying, Anil T Ahuja, and Danny Z Chen. Boxnet: Deep learning based biomedical image segmentation using boxes only annotation. *arXiv preprint arXiv:1806.00593*, 2018.
- Zhuo Zhao, Lin Yang, Hao Zheng, Ian H Guldner, Siyuan Zhang, and Danny Z Chen. Deep learning based instance segmentation in 3d biomedical images using weak annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 352–360. Springer, 2018.

# Joint Learning of Brain Lesion and Anatomy Segmentation from Heterogeneous Datasets

Nicolas Roulet<sup>1</sup>

NROULET@DC.UBA.AR

Diego Fernandez Slezak<sup>1</sup>

DFSLEZAK@DC.UBA.AR

Enzo Ferrante<sup>2</sup>

EFERRANTE@SINC.UNL.EDU.AR

<sup>1</sup> Universidad de Buenos Aires, CONICET, Buenos Aires, Argentina

<sup>2</sup> Universidad Nacional del Litoral, CONICET, Santa Fe, Argentina

## Abstract

Brain lesion and anatomy segmentation in magnetic resonance images are fundamental tasks in neuroimaging research and clinical practice. Given enough training data, convolutional neuronal networks (CNN) proved to outperform all existent techniques in both tasks independently. However, to date, little work has been done regarding simultaneous learning of brain lesion and anatomy segmentation from disjoint datasets.

In this work we focus on training a single CNN model to predict brain tissue and lesion segmentations using heterogeneous datasets labeled independently, according to only one of these tasks (a common scenario when using publicly available datasets). We show that label contradiction issues can arise in this case, and propose a novel *adaptive cross entropy* (ACE) loss function that makes such training possible. We provide quantitative evaluation in two different scenarios, benchmarking the proposed method in comparison with a multi-network approach. Our experiments suggest ACE loss enables training of single models when standard cross entropy and Dice loss functions tend to fail. Moreover, we show that it is possible to achieve competitive results when comparing with multiple networks trained for independent tasks.

**Keywords:** Brain image segmentation, heterogeneous datasets, convolutional neural networks

## 1. Introduction

Segmentation of anatomical and pathological structures in volumetric images is a fundamental task for biomedical image analysis. It constitutes the first step in several medical procedures such as shape analysis for population studies, computed assisted diagnosis/surgery and automatic radiotherapy planning, among many others. Segmentation accuracy is therefore of paramount importance in these cases, since it will necessarily influence the overall quality of such procedures.

During the last years, convolutional neural networks (CNNs) proved to be highly accurate to perform medical image segmentation (Ronneberger et al., 2015; Kamnitsas et al., 2016, 2017a; Shakeri et al., 2016). In this scenario, a training dataset consists of medical images with expert annotations associated to a particular task of interest. Following a supervised approach, CNNs are trained to perform such task by learning the network parameters that minimize a given loss function over the training data. In the context of brain image segmentation (of main interest in this work), publicly available datasets with manual annotations usually correspond to single tasks. These tasks might be associated to anatomy segmentation (e.g. brain tissues (Mendrik et al., 2015; Cocosco et al., 1997), sub-cortical structures (Rohlfing, 2012)) or pathological segmentation (e.g. brain tumours (BRATS, 2012), white matter hiper-intensities (WMH, 2017)).

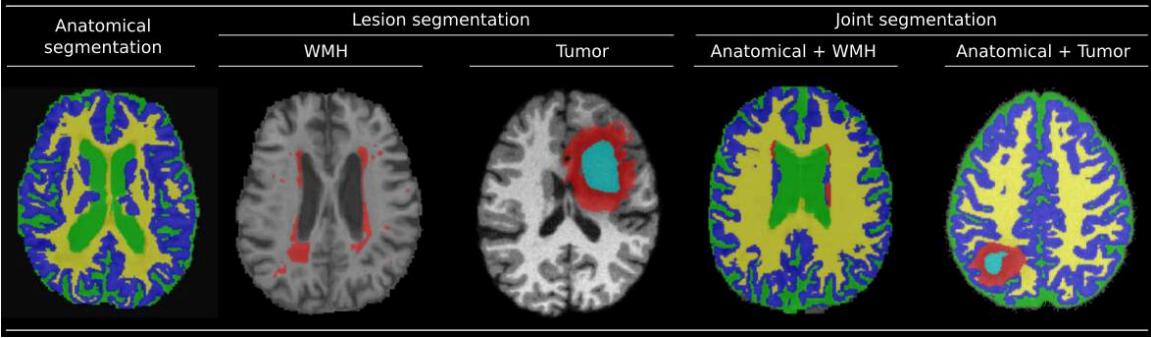


Figure 1: Example of brain MRI with overlapped annotations corresponding to anatomy, lesion and joint segmentations. Note that whatever is considered as background in both, WMH and tumor segmentation datasets, should be classified as tissue according to the anatomy dataset. This fact misleads the training process of a single CNN when using standard categorical cross entropy or Dice losses to perform joint learning of lesion and anatomy segmentation.

Even if most publicly available datasets provide image annotations for single tasks, in practice it is usually desirable to train single models which can learn to perform multiple segmentation tasks simultaneously. We focus on the particular case of brain magnetic resonance images (MRI), where segmenting both brain lesions and anatomical structures is especially relevant. For example, in the context of neurovascular and neurodegenerative diseases (Moeskops et al., 2018), white matter hyper-intensity (WMH) segmentation in brain MRI is usually combined with brain tissue segmentation when studying cognitive dysfunction in elderly patients (De Bresser et al., 2010). Another example is related to brain tumour segmentation (Menze et al., 2015). Combining brain tumor segmentation with brain tissue classification (Moon et al., 2002) would have an enormous potential value for improved medical research and biomarkers discovery. We will explore both application scenarios and provide experimental evidence about the effectiveness of the proposed method to perform joint learning of brain lesion and anatomy segmentation in these cases.

Learning to segment multiple structures from heterogeneous datasets is a challenging task, since labels coming from different datasets may contradict each other and mislead the training process. In the particular case of brain lesion and anatomy segmentation from MRI, Figure 1 illustrates this issue. Given two datasets with disjoint labels (for example, brain tissues and WMH lesions), whatever is considered as background in the lesion dataset, should be classified as tissue according to the anatomy dataset. This raises a label contradiction problem that will be studied in this work.

We interpret brain lesion and anatomy segmentation as two different tasks which are learned from heterogeneous datasets, meaning that each dataset is annotated for a single task. In what follows, we briefly describe related works about learning to segment from disjoint annotations, discuss the issues that arise when training a single CNN model to perform both tasks with standard loss functions, and propose a simple, yet effective, adaptive loss function that makes it possible to train such model using heterogeneous datasets.

### 1.1. Related Work

Similar multi-task problems in the context of image segmentation were explored in recent works. Regarding segmentation for medical images, (Moeskops et al., 2016) studied how a single deep CNN can be used to predict multiple anatomical structures for three different tasks including brain MRI, breast MRI and cardiac computed tomography angiography (CTA) segmentation. They showed that a standard combined training procedure with balanced mini-batch sampling results in segmentation performance equivalent to that of a deep CNN trained specifically for that task. This problem differs from our setting since every dataset is associated to a different organ. Therefore, labels from different datasets can not co-exists in a single image avoiding the label contradiction problem illustrated in Figure 1.

Closest to our work are those by (Fourure et al., 2017; Rajchl et al., 2018), where a single segmentation model is learned from multiple training datasets defined on images representing similar domains. In (Fourure et al., 2017), the authors train a model to perform semantic full scene labeling in outdoor images coming from different datasets with heterogeneous labels. They propose a *selective cross entropy* loss that, instead of considering a single final *softmax* activation function defined over the entire set of possible labels, is computed using a dataset-wise *softmax* activation function. This dataset-wise *softmax* only takes into account those labels available in the dataset corresponding to the current training sample. A similar strategy is followed by (Rajchl et al., 2018) in the context of brain image segmentation. The authors propose the NeuroNet, a multi-output CNN that mimics several popular and state-of-the-art brain segmentation tools producing segmentations for brain tissues, cortical and sub-cortical structures. Differently from (Fourure et al., 2017), NeuroNet combines a multi-decoder architecture (one decoder for every dataset/task) with an analogous multi-task loss based on cross entropy, defined as the average of independent loss functions computed for every single task. Note that our problem differs from those tackled in both papers: our aim is to produce a segmentation model that assigns a single label to every voxel (considering the union of anatomical and pathological labels). On the contrary, they aim at predicting one and exactly one label from each labelset for every voxel, i.e. multiple labels will be assigned to every voxel.

## 2. Learning Brain Lesion and Anatomy Segmentation from Heterogeneous Datasets

**Problem Statement:** Given a set of  $K$  heterogeneous datasets  $\{\mathcal{D}_k\}$ ,  $1 \leq k \leq K$ , let us formalize the joint learning segmentation problem. Each dataset  $\mathcal{D}_k = \{(x, y)_n\}$  is composed of pairs  $(x, y)_n$ , where  $x$  is an image and  $y$  a segmentation mask assigning a label  $l \in \mathcal{L}_k$  to every  $i$ -th voxel  $x_i$ .  $\mathcal{L}_k$  is the labelset associated to dataset  $\mathcal{D}_k$ . We assume disjoint labelsets, except for the background label included in all datasets. We aim at learning the parameters  $\Theta$  for a single segmentation model  $f(\hat{x}; \Theta)$  that, given a new image  $\hat{x}$ , produces a segmentation mask  $\hat{y}$  where every voxel  $\hat{y}_i \in \hat{\mathcal{L}} = \bigcup_{k=1}^K \mathcal{L}_k$ . The label space  $\hat{\mathcal{L}}$  is built as the union of all labelsets, and we assign a single label to every voxel  $\hat{y}_i$ .

Note that, since the new labelset  $\hat{\mathcal{L}}$  includes all labels from all datasets, some structures that were labeled as background in one dataset may be labeled as foreground in other datasets, raising the label contradiction problem shown in Figures 1 and 2.a. In these cases, the foreground labels (e.g. brain tissue labels) should prevail over the background labels in the final mask generated by the segmentation model.

In case of MRI brain lesion and anatomy segmentation, we have  $K = 2$  brain MRI datasets. The first one, denoted  $\mathcal{D}_{\mathcal{A}}$ , is annotated with anatomical (brain tissue) labels while the second one,

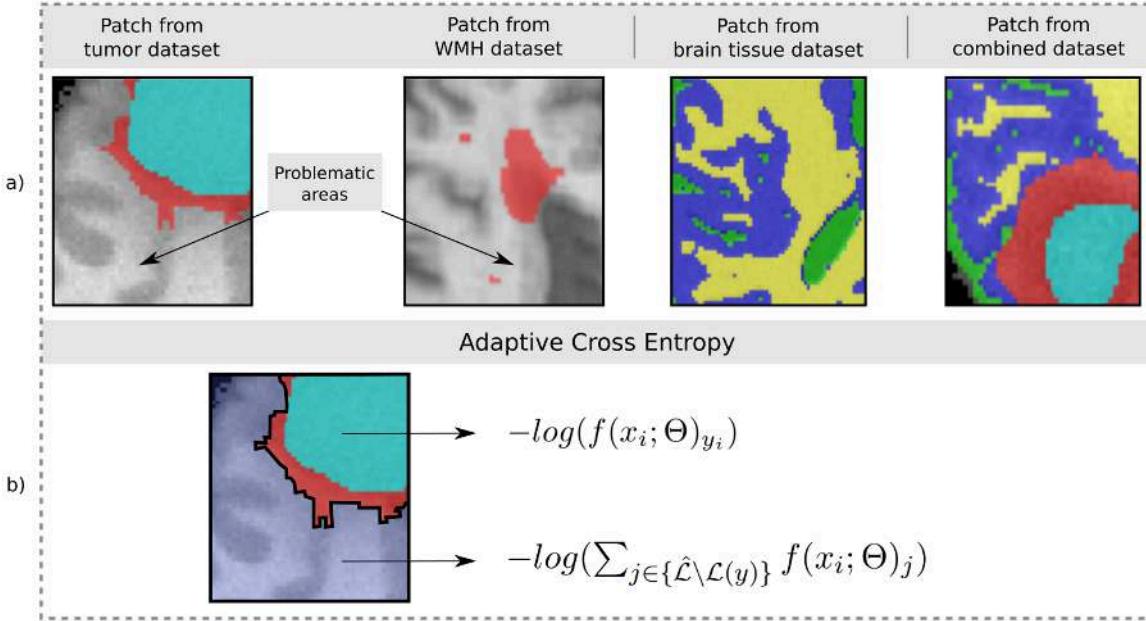


Figure 2: (a) Example of image patches with overlapped segmentation masks sampled from: the lesion datasets (tumor and WMH), the anatomical (brain tissue) dataset and the desired combined segmentation for which we do not have training data. Problematic areas are those for which the original lesion datasets indicate background label, while they should be annotated as actual tissue labels.  
 (b) The proposed *adaptive cross entropy* behaves differently depending on the structures of interest under consideration. We reinterpret the meaning assigned to the lesion background label (in blue) as '*any label that is not lesion*' and modify the loss function accordingly.

referred as  $\mathcal{D}_{\mathcal{L}}$ , considers brain lesions (tumor or WMH are the application scenarios studied in this work). The corresponding label spaces for every dataset are  $\mathcal{L}_{\mathcal{A}}$  and  $\mathcal{L}_{\mathcal{L}}$ . In what follows, we describe multiple alternatives to train such model based on a standard U-Net architecture ([Ronneberger et al., 2015](#)).

## 2.1. Naive Models

We first consider a naive model where a single U-Net is trained by minimizing standard loss functions (typical categorical cross entropy and Dice losses), to perform joint learning from heterogeneous datasets. We employ a standard U-Net architecture (see Appendix A for a complete description of the architecture) with a final *softmax* layer producing  $|\hat{\mathcal{L}}|$  probability maps, i.e. one for each class in the joint labelset  $\hat{\mathcal{L}}$ . Patch-based training is performed by constructing balanced mini-batches of image patches. We balance the mini-batches by sampling with equal probability from all datasets and all classes.

As stated in section 1.1 and illustrated in Figure 2.a, labels coming from different datasets may contradict each other and mislead the training process. Brain tissue segmentations or cortical/sub-

cortical structures generally cover the complete brain mass. However, lesion annotations like WMH and tumour cover only a small portion of it. The main issue with the proposed naive model arises from this fact: when sampling image patches containing small lesions, whatever is considered background in the patch should be actually classified as some type of brain tissue. However, since the lesion dataset does not contain brain tissue annotations, it will be considered as background. In other words, the model will be encouraged to classify brain tissue as background. In the results that will be presented in Section 3, we provide empirical evidence of this issue and its impact in model performance.

## 2.2. Multi-network Baseline

A trivial solution to the aforementioned problem is to use multiple independent models, trained for every specific task. In this case, segmentation results are then combined following some kind of fusion scheme. In case of brain lesion and tissue segmentation, since lesion labels prevail over tissue labels, we can simply overwrite them. However, note that such model requires extra efforts at training time: we need to train a single model for every dataset, increasing not only the training time but also the overall model complexity, i.e. the number of learned parameters. Moreover, at test time, every model is evaluated on the test image and a label fusion strategy must be applied to combine the multiple predictions.

We consider a multi U-Net model as baseline to benchmark the proposed solution, training a single U-Net with categorical cross entropy in every dataset. Label fusion is implemented by overwriting the brain tissue segmentation with the (non-background) lesion masks.

## 2.3. Adaptive Cross Entropy

In this work, we propose to overcome the issues that arise when training a single CNN from heterogeneous (and potentially contradictory) datasets with a new loss function titled *adaptive cross entropy* (ACE). Let us first recall the classical formulation of cross entropy. Given an estimate distribution  $q$  for a true probability distribution  $p$  defined over the same discrete set (in our setting, the set  $\hat{\mathcal{L}}$  of possible labels, with  $C = |\hat{\mathcal{L}}|$ ), the cross entropy between them is computed as:

$$H(p, q) = - \sum_{j=1}^C p_j \cdot \log(q_j). \quad (1)$$

For a given voxel  $x_i$  with ground-truth label  $y_i \in \hat{\mathcal{L}}$  (with  $1 \leq y_i \leq C = |\hat{\mathcal{L}}|$ ), we compute the categorical cross entropy loss between the voxel-wise model prediction  $f(x_i; \Theta)$ , and the corresponding one-hot encoded version of  $y_i$  denoted by  $e^{(y_i)}$  as:

$$\begin{aligned} H(x_i, y_i) &= - \sum_{j=1}^C e_j^{(y_i)} \cdot \log(f(x_i; \Theta)_j) = - \sum_{j=1}^C \mathbb{1}_{[y_i=j]} \cdot \log(f(x_i; \Theta)_j) \\ &= -\log(f(x_i; \Theta)_{y_i}). \end{aligned} \quad (2)$$

The standard voxel-wise cross entropy loss  $L_H$  is aggregated as the average loss considering all voxels  $\{x_i\}_{1 \leq i \leq m}$  in the image patch:

$$L_H(x, y) = - \sum_{i=1}^m \log(f(x_i; \Theta)_{y_i}). \quad (3)$$

The cross entropy loss  $L_H$  is minimized when the prediction equals the ground-truth. In the multi-task context discussed in this work, this raises the label contradiction problem between lesion background and brain tissue segmentation illustrated in Figure 2.a. This fact motivates the design of the *adaptive cross entropy* (ACE) loss which behaves differently depending on the structures of interest under consideration. We reinterpret the meaning assigned to the background label of the lesion dataset as ‘any label that is not lesion’ and modify the loss function accordingly. The proposed *adaptive cross entropy* is therefore defined as:

$$H^A(x_i, y_i) = \begin{cases} -\log(f(x_i; \Theta)_{y_i}) & \text{if } y_i \text{ is not lesion background} \\ -\log(\sum_{j \in \{\hat{\mathcal{L}} \setminus \mathcal{L}(y)\}} f(x_i; \Theta)_j) & \text{if } y_i \text{ is lesion background} \end{cases} \quad (4)$$

where the set  $\{\hat{\mathcal{L}} \setminus \mathcal{L}(y)\}$  contains all labels, except those in the current image patch ground-truth (referred as  $\mathcal{L}(y)$ ). Equation 4 shows that ACE employs the standard cross entropy formulation when voxel  $i$  is labeled as anything but lesion background. However, when voxel  $i$  corresponds to lesion background, we compute  $-\log(s)$ , where  $s = \sum_{j \in \{\hat{\mathcal{L}} \setminus \mathcal{L}(y)\}} f(x_i; \Theta)_j$  is the sum of scores  $f(x_i; \Theta)_j$  for all classes  $j$  that are not present in the patch  $y$  (including background). In this way, when the label is not in conflict, minimizing  $H^A$  is equivalent to maximizing the score for the correct class. However, when dealing with a voxel whose ground truth is lesion background (i.e. we are not sure about the brain tissue that corresponds to it), the model tends to maximize the probability for all non-lesion classes. Figure 2.b illustrates this idea. In practice, we compute the aggregated ACE loss  $L_H^A$  for all voxels  $\{x_i\}_{1 \leq i \leq m}$  in the image patch as:  $L_H^A(x, y) = \frac{1}{m} \sum_{i=1}^m H^A(x_i, y_i)$ .

Note that in the ACE formulation, we sum over the scores before taking the logarithm. The reasoning behind having the sum inside the log function on the proposed adaptive cross entropy is to effectively unify those labels that are not lesion (i.e. background and brain tissue segmentations, which raise the label contradiction problem illustrated in Figure 2.a) in a unique class. We do that by assigning to this virtual class the sum of the scores the model assigned to each of those labels.

Note that in the application scenarios studied in this work, lesion labels collide with brain tissues, motivating the ACE formulation given in Equation 4. Nonetheless, given an arbitrary number of  $K$  datasets, in general it is straightforward to apply the proposed ACE loss to different labels raising similar issues, by just changing the condition that adapts the loss behaviour.

### 3. Experiments & Results

Six different datasets were used in the experimental comparative analysis. We consider joint learning of brain tissue segmentation and two separate type of lesions: brain tumor and WMH. We trained models specialized for brain tissue + WMH, and other models for brain tissue + tumor, showing that the proposed ACE loss function can generalize to different scenarios.

#### BRAIN TISSUES + WMH SCENARIO

We employed the training data provided by the MRBrainS13 Challenge (Mendrik et al., 2015) (brain tissue annotations), the WMH Segmentation Challenge (WMH, 2017) (WMH lesions) and MRBrains18 (MRBrainS, 2018) (brain tissues + WMH). We trained/validated our models using the

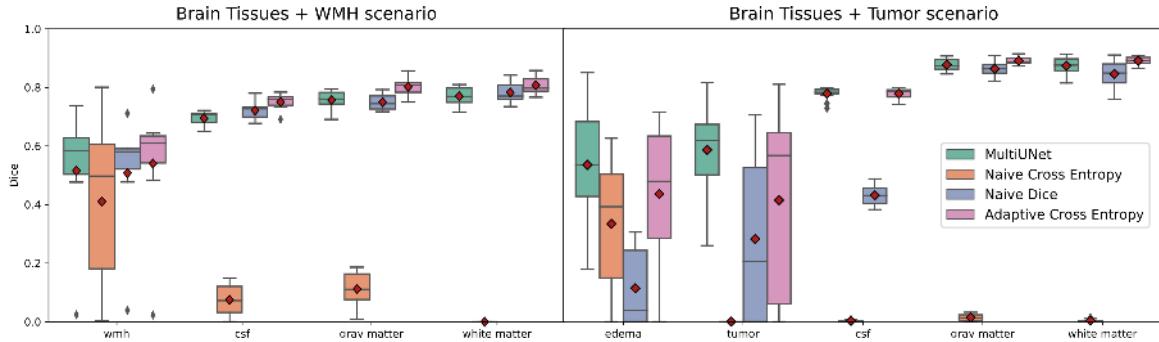


Figure 3: Experimental results obtained when comparing a single model trained with the proposed ACE loss, with the Multi-UNet and the naive cross entropy and Dice models (red diamond indicates the mean value). Note that a single model trained with ACE achieves equivalent performance to that of Multi-UNet, while naive models under-perform by a big margin in both cases.

Table 1: Numerical results corresponding to the experiments shown in Figure 3.

	Brain Tissues + WMH										Brain Tissues + Tumor									
	WMH		CSF		GM		WM		Edema		Tumor		CSF		GM		WM			
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Multi UNet	0.516	0.232	0.694	0.028	0.757	0.035	0.77	0.035	0.509	0.228	0.586	0.143	0.778	0.021	0.877	0.02	0.874	0.026		
Naive CE	0.411	0.294	0.075	0.057	0.112	0.067	0	0	0.335	0.219	0	0	0.002	0.003	0.013	0.011	0.003	0.004		
Naive Dice	0.508	0.218	0.721	0.035	0.75	0.029	0.783	0.038	0.114	0.126	0.282	0.252	0.432	0.032	0.863	0.025	0.846	0.042		
ACE	0.54	0.245	0.75	0.031	0.802	0.034	0.807	0.033	0.414	0.264	0.415	0.3	0.779	0.018	0.891	0.013	0.891	0.012		

training partition of MRBrainS13 as anatomical dataset ( $\mathcal{D}_A$ ) and WMH Segmentation Challenge as lesion dataset ( $\mathcal{D}_L$ ). For testing, we used the joint segmentations provided for training in the MR-BrainS2018 Challenge, to evaluate the simultaneous predictions. The data from the MRBrainS13 Challenge consists of 5 images with brain tissue annotations, of which 4 were used for training, and the remaining one for validation. The WMH Segmentation Challenge provides 60 images with the corresponding WMH reference segmentation, of which 48 were used for training, and the rest for validation. The MRBrainS18 Challenge provides 7 images, which were all used for evaluation.

#### BRAIN TISSUES + TUMOR SCENARIO

Given the lack of datasets with simultaneous annotations for brain tumors and tissues, we resorted to using synthetic and simulated images. We trained/validated our models using 15 images from the Brainweb (Cocosco et al., 1997) synthetic brain phantoms with brain tissue annotations for the anatomical dataset ( $\mathcal{D}_A$ ). For the lesion dataset ( $\mathcal{D}_L$ ) we employed 50 simulated tumor images available from the BRATS2012 challenge (BRATS, 2012). For testing, we simulated 20 brain tumors using Tumorsim (Prastawa et al., 2009), using 5 healthy Brainweb phantom probability maps. In that way, combined segmentations of brain tissue and tumors were available for testing. Note that, for the sake of fairness, healthy images used to simulate brain tumors for testing were not included in the training dataset ( $\mathcal{D}_A$ ).

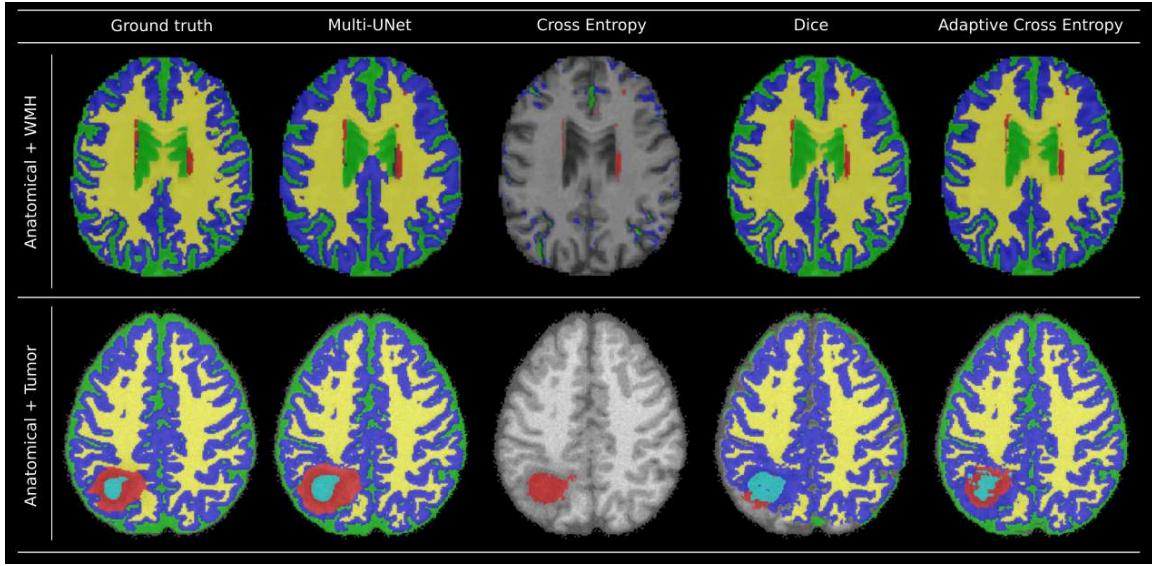


Figure 4: Qualitative results for both scenarios (brain tissues + WMH in the top row, and brain tissues + tumor segmentation in the bottom row). Note that using naive cross entropy and Dice losses result in very poor performance. The proposed ACE makes it possible to train a single model for both tasks with equivalent performance to multiple networks by solving the label contradiction issues.

## RESULTS & DISCUSSION

Figure 3 summarizes the quantitative results for both application scenarios, when comparing the Multi-UNet model with single models trained with naive cross entropy and Dice functions as well as the proposed ACE<sup>1</sup> (see Figure 4 for qualitative results). As expected, the Multi-UNet model trained with standard cross entropy outperforms the single models trained with naive losses. More importantly, our proposed ACE makes it possible to train a single model for joint learning of brain lesion and anatomy from heterogeneous datasets, achieving equivalent performance to that of Multi-UNet.

This is due to the fact that both, Multi-UNet and the single ACE models, are not affected by the label contradiction problem illustrated in Figure 2.a. Note that in case of brain tissue segmentation, the single model trained with ACE tends to outperform even the Multi-UNet model. As discussed in (Rajchl et al., 2018), learning jointly from hierarchical sets of class labels has the potential to increase the overall accuracy based on theory derived from multi-task learning. We hypothesize that this increase in performance is related to this fact: since the model trained with ACE learns to predict lesion and tissues simultaneously, it can also learn label interactions that the Multi-UNet can not capture.

1. We implemented the CNN in Keras and trained it using Adam optimizer with default parameters. Balanced mini-batches of 7 image patches of size  $32 \times 32 \times 32$  are used during training. A complete description of the baseline UNet architecture used for both, single and multi-network models, is provided in Appendix A.

A deeper analysis of the quantitative results reveals that the single UNet model trained with the proposed ACE achieved equivalent performance to the Multi-UNet in WMH segmentation (no significant differences according to Wilcoxon test), better or equivalent performance in terms of brain tissue segmentation (depending on the brain structure) and only worse performance for edema and tumor. This worse performance for edema and tumor is explained by the fact that the Multi-UNet was trained using all available modalities per dataset, while the single UNet was trained using only those modalities available in both, anatomical and lesion datasets. This is a limitation of our approach when compared with multiple UNets trained for specific tasks: since we perform joint training of a single model with fixed number of input channels, we can only use those sequences available in both anatomy and lesion datasets. In case of edema and brain tumor segmentation, the Multi-UNet was trained with multiple MR modalities for the tumor segmentation task (it uses T1, T1g, T2 and FLAIR) while the single UNet was trained using only T1 images (all details about available MR modalities for every dataset are provided in Appendix B). This requirement may represent a limitation if the datasets depend on different types of image modalities. There are alternatives that could be considered to deal with this issue like imputing the missing modalities by means of image synthesis or using ad-hoc techniques like the HeMIS (Hetero-Modal Image Segmentation) model by (Havaei et al., 2016).

Even if all images used in the experiments are MRI, there is a shift in the distribution of image intensities when we go from datasets used at training and test time. This is known as the multi-domain problem, and is usually addressed using domain adaptation techniques (Kamnitsas et al., 2017b). In this work, we did not take into account the multi-domain problem. In the future, we plan to extend the proposed method and incorporate domain adaptation, further improving the accuracy of the results.

#### 4. Conclusions

In this work we proposed the *adaptive cross entropy* loss, a novel function to perform joint learning of brain lesion and anatomy segmentation from heterogeneous datasets using CNNs. The proposed loss takes into account potential label contradiction conflicts that can arise when training segmentation algorithms for multiple tasks using datasets with disjoint annotations. We trained single CNN models using the proposed ACE, naive cross entropy and Dice losses, and compared their performance with a Multi-UNet model where independent CNNs were trained for every task. Experimental evaluations in two scenarios provided empirical evidence about the effectiveness of the proposed approach.

In the future, we plan to extend the evaluation of the proposed loss function to other CNN architectures (Deepmedic (Kamnitsas et al., 2016) for example) and to alternative brain MRI segmentation scenarios (e.g. considering subcortical structures as anatomical segmentation or traumatic brain injuries as lesions). Moreover, we plan to investigate the effects of the multi-domain problem in this context, and incorporate domain adaptation strategies to address this issue when learning from heterogeneous datasets.

Regarding the ACE formulation, we plan to explore alternative weighting mechanisms within the loss function that could help to alleviate the class-imbalance problems that could emerge when dealing with tiny structures of interest.

## Acknowledgments

NR is now at Google. EF is beneficiary of an AXA Research Grant. We thank NVIDIA Corporation for the donation of the Titan X GPU used for this project. DFS is partially supported by Universidad de Buenos Aires and CONICET.

## References

- BRATS. MICCAI 2012 Challenge on Multimodal Brain Tumor Segmentation. <http://www.imm.dtu.dk/projects/BRATS2012>, 2012. [Online].
- Chris A Cocosco, Vasken Kollokian, Remi K-S Kwan, G Bruce Pike, and Alan C Evans. Brainweb: Online interface to a 3d mri simulated brain database. In *NeuroImage*. Citeseer, 1997.
- Jeroen De Bresser, Audrey M Tiehuis, Esther Van Den Berg, Yael D Reijmer, Cynthia Jongen, L Jaap Kappelle, Willem P Mali, Max A Viergever, Geert Jan Biessels, Utrecht Diabetic Encephalopathy Study Group, et al. Progression of cerebral atrophy and white matter hyperintensities in patients with type 2 diabetes. *Diabetes care*, 2010.
- Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Natalia Neverova, Alain Tréneau, and Christian Wolf. Multi-task, multi-domain learning: Application to semantic segmentation and pose regression. *Neurocomputing*, 251:68 – 80, 2017. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2017.04.014>. URL <http://www.sciencedirect.com/science/article/pii/S0925231217306847>.
- Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Heteromodal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477. Springer, 2016.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Deepmedic for brain tumor segmentation. In *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 138–149. Springer, 2016.
- Konstantinos Kamnitsas, Wenjia Bai, Enzo Ferrante, Steven McDonagh, Matthew Sinclair, Nick Pawlowski, Martin Rajchl, Matthew Lee, Bernhard Kainz, Daniel Rueckert, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. In *International MICCAI Brainlesion Workshop*, pages 450–462. Springer, 2017a.
- Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging*, pages 597–609. Springer, 2017b.

Adriënne M Mendrik, Koen L Vincken, Hugo J Kuijf, Marcel Breeuwer, Willem H Bouvy, Jeroen De Bresser, Amir Alansary, Marleen De Bruijne, Aaron Carass, Ayman El-Baz, et al. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. *Computational intelligence and neuroscience*, 2015:1, 2015.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993, 2015.

Pim Moeskops, Jelmer M Wolterink, Bas HM van der Velden, Kenneth GA Gilhuijs, Tim Leiner, Max A Viergever, and Ivana Išgum. Deep learning for multi-task medical image segmentation in multiple modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 478–486. Springer, 2016.

Pim Moeskops, Jeroen de Bresser, Hugo J Kuijf, Adriënne M Mendrik, Geert Jan Biessels, Josien PW Pluim, and Ivana Išgum. Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in mri. *NeuroImage: Clinical*, 17:251–262, 2018.

Nathan Moon, Elizabeth Bullitt, Koen Van Leemput, and Guido Gerig. Automatic brain and tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 372–379. Springer, 2002.

MRBrainS. MRBrainS18. <http://mrbrains18.isi.uu.nl/>, 2018. [Online].

Marcel Prastawa, Elizabeth Bullitt, and Guido Gerig. Simulation of brain tumors in mr images for evaluation of segmentation efficacy. *Medical image analysis*, 13(2):297–311, 2009.

Martin Rajchl, Nick Pawlowski, Daniel Rueckert, Paul M Matthews, and Ben Glocker. Neuronet: Fast and robust reproduction of multiple brain image segmentation pipelines. *International Conference on Medical Imaging with Deep Learning (MIDL) 2018*, 2018.

Torsten Rohlfing. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE transactions on medical imaging*, 31(2):153–163, 2012.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL <http://arxiv.org/abs/1505.04597>.

Mahsa Shakeri, Stavros Tsogkas, Enzo Ferrante, Sarah Lippe, Samuel Kadoury, Nikos Paragios, and Iasonas Kokkinos. Sub-cortical brain structure segmentation using f-cnn’s. *ISBI 2016*, 2016.

WMH. WMH Segmentation Challenge. <http://wmh.isi.uu.nl/>, 2017. [Online].

## Appendix A. Detailed Network Architecture

The architecture used in this work is based on a standard U-Net ([Ronneberger et al., 2015](#)). It can be divided into a contraction and an expansion path. Each path is a sequence of four convolution blocks, composed of two convolutional layers with  $3 \times 3 \times 3$  kernels and one voxel of padding, each one followed a ReLU activation layer. We also used batch-normalization to ease training. Every block from the contraction path is connected to the next one by a  $2 \times 2 \times 2$  max-pooling layer, while the blocks from the expansion path are connected by  $2 \times 2 \times 2$  transposed convolutions for upsampling. The output from each block of the contraction path is added to the input of the corresponding block from the expansion path to combine the localized features of the former with the high level information from the latter. This is in contrast with standard U-Net which uses concatenation of feature maps instead of summation. The layers from the first block have 32 channels. The number of channels is doubled in every max-pooling layer and halved in every transposed convolution layer. Finally, a  $1 \times 1 \times 1$  convolution layer with softmax activation is used to convert the output of the last layer into voxel-wise label probability maps.

We implemented the CNN in Tensorflow and trained it using Adam optimizer. The weights were initialized using He method ([He et al., 2015](#)). Balanced mini-batches of 7 image patches of size  $32 \times 32 \times 32$  were used during training.

## Appendix B. MR Sequences Available Per Dataset

Different MR sequences were available for every dataset. Table 2 summarizes this information.

Table 2: MR sequences available per dataset.

Scenario	Dataset	T1	T1 with Gadolinium (T1g)	T2	IR	FLAIR
Brain Tissue + WMH	MRBrains13	X			X	X
	WMH	X				X
	MRBrains18	X			X	X
Brain Tissue + Tumor	BrainWeb	X				
	BRATS12	X	X	X		X
	Tumorsim	X	X	X		X

The UNet architecture used in our experiments can receive multiple MR sequences as input by simply interpreting them as multiple image channels. Note that the Multi-UNet network was trained with as many sequences as possible per task. For example, if T1, T2 and FLAIR sequences were available in the lesion dataset and only T1, T2 were available for the anatomy dataset, we trained every independent UNet using all available sequences (of course, these sequences have to be available in the test dataset as well). However, when training the single UNet models using the naive losses and ACE, we can only use those sequences available in both anatomy and lesion datasets.

Given the MR sequences available for every dataset (shown in Table 2) we trained the single and multi-network models under the following setting:

- Brain Tissue + WMH scenario: The Multi-UNet model was trained and tested using T1+IR+FLAIR for the brain tissue segmentation task, and T1+FLAIR for the WMH segmentation task. The single UNet models were trained using only T1+FLAIR for all tasks.

- Brain Tissue + Tumor scenario: The Multi-UNet model was trained and tested using T1 for the brain tissue segmentation task, and T1+T1g+T2+FLAIR for the tumor segmentation task. The single UNet models were trained using only T1 for all tasks.

Note that this setting gives some advantages to the Multi-UNet model over the single model trained with ACE, since it uses more MR sequences for the lesion segmentation task. This is reflected in the results shown in Figure 3, specially for the brain lesion segmentation task, where the better performance shown by the Multi-UNet model with respecto to the single model trained with ACE can be explained by this difference in the number of sequences used to train them.

# Learning with Multitask Adversaries using Weakly Labelled Data for Semantic Segmentation in Retinal Images

**Oindrila Saha**

**Rachana Sathish**

**Debdoott Sheet**

*Indian Institute of Technology Kharagpur*

OINDRILA\_SAHA13@IITKGP.AC.IN

RACHANA.SATHISH@IITKGP.AC.IN

DEBDOOT@EE.IITKGP.AC.IN

## Abstract

A prime challenge in building data driven inference models is the unavailability of statistically significant amount of labelled data. Since datasets are typically designed for a specific purpose, and accordingly are weakly labelled for only a single class instead of being exhaustively annotated. Despite there being multiple datasets which cumulatively represents a large corpus, their weak labelling poses challenge for direct use. In case of retinal images, they have inspired development of data driven learning based algorithms for segmenting anatomical landmarks like vessels and optic disc as well as pathologies like microaneurysms, hemorrhages, hard exudates and soft exudates. The aspiration is to learn to segment all such classes using only a single fully convolutional neural network (FCN), while the challenge being that there is no single training dataset with all classes annotated. We solve this problem by training a single network using separate weakly labelled datasets. Essentially we use an adversarial learning approach in addition to the classically employed objective of distortion loss minimization for semantic segmentation using FCN, where the objectives of discriminators are to learn to (a) predict which of the classes are actually present in the input fundus image, and (b) distinguish between manual annotations vs. segmented results for each of the classes. The first discriminator works to enforce the network to segment those classes which are present in the fundus image although may not have been annotated i.e. all retinal images have vessels while pathology datasets may not have annotated them in the dataset. The second discriminator contributes to making the segmentation result as realistic as possible. We experimentally demonstrate using weakly labelled datasets of DRIVE containing only annotations of vessels and IDRIID containing annotations for lesions and optic disc. Our method using a single FCN achieves competitive results over prior art for either vessel or optic disk or pathology segmentation on these datasets.

**Keywords:** Adversarial learning, convolutional neural networks, multitask learning, semantic segmentation, retinal image analysis.

## 1. Introduction

The precise segmentation of retinal anatomies and pathologies serves as an important tool for diagnosis and evaluation of various metabolic and ophthalmic disorders such as diabetes, hypertension, glaucoma and choroidal neovascularization, etc. (Bowling, 2015). Experts can analyze changes in vascular morphology by segmenting the retinal vessels. Vessel and optic disk segmentation followed by localization of fovea region by Haddouche et al. (2010) is the first step for the quantitative analysis of retinal images which is helpful for diabetic retinopathy (DR) screening and diagnosis (Teng et al., 2002). DR is the leading cause of blindness in the working-age population. Screening for DR and monitoring disease progression, especially in the early asymptomatic stages, is effective

for preventing visual loss and reducing costs for health systems ([Nentwich and Ulbig, 2015](#)). The most common signs of DR are red lesions symptomatic of microaneurysms, hemorrhages and bright lesions symptomatic of exudates. However, it is tedious and time consuming to segment retinal diseases or anatomies manually, especially in fundus images. Thus ability to automatically and reliably segment all structures is a long-standing problem since decades.

One of the major challenges in this task is the absence of a single dataset which contains exhaustive pixel level semantic annotation of all parts of the retina. However this task of creating such a dataset being a highly taxing job, we propose to employ multiple readily available datasets which have reliable annotations for some of the classes, with no necessity of any single dataset having all given classes annotated and there also not being common classes annotated across different datasets. This paper proposes a single model that can learn from separate datasets to predict all classes in a given retinal image. We address this problem by utilizing multiple adversaries for addressing it as a multitask approach. The convolutional neural network for semantic segmentation tries to generate segmentation maps which visually resemble the ground truth. Two discriminators are used; one for distinguishing between manual vs. segmented maps and the other for determining which of the various classes are present in the fundus image. The first discriminator contributes to making the segmentation result as realistic as possible, thus learning to capture the finer details. The second discriminator works to enforce the model to learn to segment for all classes which are present and are not just limited to those which are annotated in a particular image. Results show the superior performance of the proposed method over several existing methods of vessel segmentation in terms of sensitivity, specificity and classification accuracy. The method reduces over-segmentation i.e. the presence of false positives as compared to previous state-of-the-art. Comparison with the results from the leaderboard in the Diabetic Retinopathy Segmentation and grading challenge <sup>1</sup> show that proposed method achieves better area under Precision vs Sensitivity curve for the different lesions and surpasses Jaccard Index scores for Optic Disk segmentation.

The paper is organized as follows. The existing methods for vessel and lesion segmentation are analyzed in Section 2. A thorough description of the overall proposed adversarial learning method and network architectures is presented in Section 3. Section 4 presents details of the experiments, setup and evaluation criteria. In Section 5, the results are presented and compared to existing methods. Section 6 concludes the paper.

## 2. Related Work

The existing segmentation methods can be divided into supervised or unsupervised categories according to whether the manual labeled ground truths are used or not.

[Azzopardi and Petkov \(2013\)](#) proposed to use a matched filtering method which convolves a 2-D kernel with the retinal image where the matched filter response indicates the presence of the feature. [Salazar-Gonzalez et al. \(2014\)](#) first carried out a preprocessing for the image by adaptive histogram equalization and robust distance transform, and then segmented retinal vessels with graph cut. Other methods include the vessel profile analysis ([Wang et al., 2007](#)), active contour based segmentation ([Al-Diri et al., 2009](#)), line detection ([Ricci and Perfetti, 2007](#)) and cluster based models ([Emary et al., 2014](#)). In case of diabetic retinopathy the first solutions were trained to detect lesions using manual segmentation for supervised classification ([Mookiah et al., 2013](#)). [Ali et al. \(2013\)](#) uses a statistical atlas for exudate segmentation, while [Inoue et al. \(2013\)](#) uses Eigen value analysis and

---

1. <https://idrid.grand-challenge.org/Home/>

Hessian matrix for microaneurysm segmentation. Most of these models employed a supervised approach for the specific segmentation task and employed curated datasets suitable only for a given task in hand, weakly labelling only a few of all the available classes in the images.

Supervised learning generally requires annotated data in order to build a predictive model. These methods can be regarded as a pixel-level binary classification problem. Each pixel belongs to vessel or non-vessel. There are two parts to segmentation using supervised approaches. One is an extractor to extract the feature vectors of pixels; the other one is a classifier to map extracted vectors to the corresponding labels for each pixel. A number of feature extractors like the Gabor filters (Akram et al., 2014), the Gaussian filter (Zhang et al., 2010) have been used previously. Various classifiers such as k-NN classifier (Staal et al., 2004), support vector machine (SVM) (Gandhi and Dhanasekaran, 2013), artificial neural networks (ANN) (García et al., 2009), AdaBoost (Lupascu et al., 2010) etc, have been proposed to deal with the task.

With the advent of deep learning, there have been numerous studies that investigated the vessel segmentation problems using convolutional neural networks (CNNs). Liskowski and Krawiec (2016) proposed a supervised segmentation architecture that used a deep neural network learned with a large training dataset which was preprocessed via global contrast normalization, zero-phase whitening, geometric transformations and gamma corrections. Tan et al. (2017) used a CNN for segmentation of haemorrhages, exudates and microaneurysms. Fu et al. (2016) regarded the segmentation as a boundary detection problem and they combined the CNN and conditional random field (CRF) layers into an integrated deep network to achieve their goal.

The effectiveness of the method proposed by Luc et al. (2016) and Ganin et al. (2016) for semantic segmentation tasks gives rise to the motivation for using a adversarial approach. The next section describes in detail the proposed multitask adversarial learning method for retinal anatomy and pathology segmentation while learning the single task from multiple weakly labelled datasets.

### 3. Methodology

In the proposed approach, the challenge of segmentation is formulated as a task of pixel-wise classification of the retinal images using a fully-convolutional neural network. An adversarial approach is used for training of the segmentation network as shown in Figure 1. Given an image  $\mathcal{I}$  of size  $3 \times M \times N$  the problem of segmentation can be formulated as a pixel-wise classification task where each pixel is assigned a label  $\mathcal{L} \in \{l_1, l_2, \dots, l_C\}$  such that an output tensor  $\mathbf{O}$  of size  $C \times M \times N$  is generated, where  $C$  is the number of classes. Each channel of  $\mathbf{O}$  corresponds to one of the classes: *background, retinal vessels, optic disk, microaneurysms, hemorrhages, hard exudates and soft exudates*. The background is excluded for calculation of segmentation loss and for input to the discriminators, so as to prevent inconsistency in the form of the model learning optic disk as a part of background for DRIVE (Staal et al., 2004) images and retinal vessels as background for IDRiD<sup>2</sup> images. The input to *Discriminator 1* is the retinal image concatenated with the channels of the output segmentation map without the background channel. These channels are randomly shuffled by a *ChannelShuffler()*. It provides a one-hot vector  $\hat{\mathbf{n}}_1$  of length  $C - 1$  in the output which determines which of the channels were present in the input retinal image. In case of a healthy retinal image which consists of only retinal vessels and optic disk without any pathology, the target vector would be  $\mathbf{n}_1 = \{1, 1, 0, 0, 0, 0\}$ , while if in a given image the only pathology present is hemorrhage, then  $\mathbf{n}_1 = \{1, 1, 0, 1, 0, 0\}$ .

---

2. <http://dx.doi.org/10.21227/H25W98>

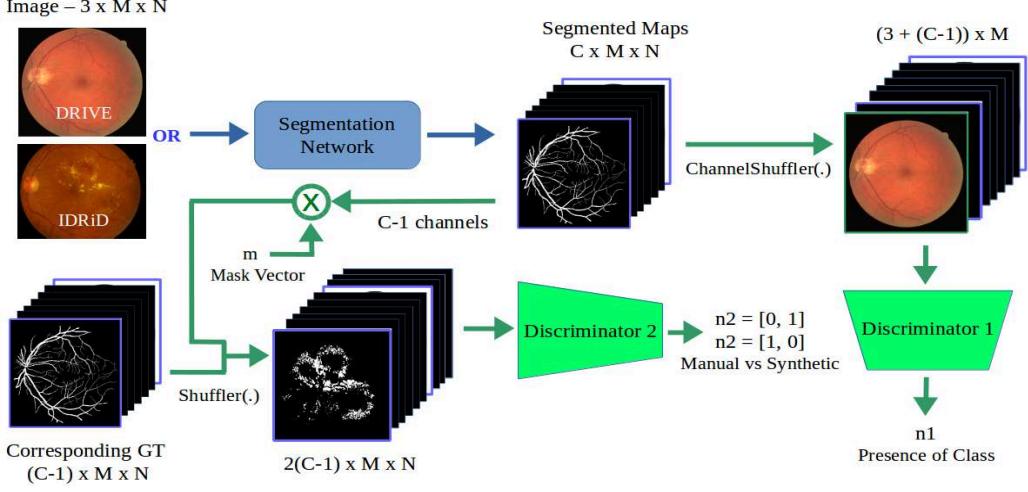


Figure 1: The proposed methodology: Segmentation network outputs a  $C$  channel tensor. Discriminator 1 is fed the shuffled segmented maps concatenated with the original image. The segmented maps are masked with vector  $\mathbf{m}$  and then concatenated with the groundtruth maps in random order and fed to Discriminator 2.

The task of *Discriminator 2* is to distinguish between manual and segmented maps. The segmented maps are masked with a one-hot vector  $\mathbf{m} = \{m_1, m_2, \dots, m_C\}$ , where  $m_c$  is 1 if the class is annotated and 0 otherwise. This masking is done so that when the ground truth of a class is not present i.e. blank, the output segmented channel of that class is also masked out to avoid uncertainty in loss measures. Thus, *Discriminator 2* will focus only on differentiating between manual vs. segmented maps for the classes that are annotated in a particular image. The manual and segmented maps are concatenated in random order with probability  $p$  from an uniform distribution  $p \sim U(0, 1)$  and fed in one go. *Discriminator 2* provides as a output a one-hot vector  $\hat{\mathbf{n}}_2$  of length 2 for this task. Discriminators are trained using Binary Cross Entropy loss function given as

$$L_{D_i}(\hat{\mathbf{n}}_i, \mathbf{n}_i) = -\frac{1}{C_{D_i}} \sum_{j=1}^{C_{D_i}} (n_{i,j} \log(\hat{n}_{i,j}) + (1 - n_{i,j}) \log(1 - \hat{n}_{i,j})) \quad (1)$$

where  $\mathbf{n}_i = \{n_{i,j}\}$ ,  $\hat{\mathbf{n}}_i = \{\hat{n}_{i,j}\}$  are the target and output for the discriminator  $D_i$  respectively for  $i \in 1, 2$  and  $C_{D_i}$  denotes the length of the output vector of  $D_i$ . *Discriminator 1* corresponds to  $D_1$  with  $C_{D_1} = 6$  and *Discriminator 2* corresponds to  $D_2$  with  $C_{D_2} = 2$ .

The basic segmentation loss is also calculated using Binary Cross Entropy loss. During training of the segmentation network a weighted sum of the segmentation loss and discriminator loss is used.

$$L_{\text{Seg-BCE}}(\mathbf{O}, \mathbf{T}) = -\frac{1}{MN} \sum_{k=1}^{C-1} \sum_{i=1}^M \sum_{j=1}^N (T_{i,j,k} \log(O_{i,j,k}) + (1 - T_{i,j,k}) \log(1 - O_{i,j,k})) \quad (2)$$

$$L_{\text{Seg}}(\mathbf{O}, \mathbf{T}) = \lambda L_{\text{Seg-BCE}}(\mathbf{O}, \mathbf{T}) + L_{D_1}(\hat{\mathbf{n}}_1, \mathbf{n}_1) + L_{D_2}(\hat{\mathbf{n}}_2, \mathbf{n}_2) \quad (3)$$

where the trade-off coefficient  $\lambda$  balances two objective functions.  $\mathbf{T} = \{T_{i,j,k}\}$ ,  $\mathbf{O}$  are the target and output for the segmentation network respectively and  $C$  denotes the number of classes.

### 3.1. Dataset Normalization

Retinal images have irregularities in illumination and noise, thus enhancements are performed on the fundus images. On every input image  $I$ , contrast limited adaptive histogram equalization is applied to enhance the contrast and also ensure that the images of both the datasets become more similar to each other, for the model to be able to predict classes which are not annotated. A  $3 \times 3$  median filter is then applied to reduce the noise in background of the image. Finally the intensity values are scaled to  $[0, 1]$  to obtain the preprocessed image  $\mathcal{I}$  to be operated upon.

### 3.2. Network Structure

Inspired by the fully convolutional networks (Long et al., 2015) a CNN architecture with skip connections is used as the segmentation network. The main path follows the typical architecture of the VGG 16 network (Simonyan and Zisserman, 2015). It consists of the repeated application of  $3 \times 3$  convolutions with padding for same size, each followed by a rectified linear unit and a  $2 \times 2$  max pooling operation with 2 pixels stride for downsampling to reduce the amount of parameters and computation. At each downsampling step the number of feature channels are doubled. In order to concatenate different feature maps through the skip connection, an upsampling of the feature map followed by a  $2 \times 2$  deconvolution (Long et al., 2015) is used, which is initialized with bilinear interpolation filters. This skip-connection is crucial so as to propagate context information to higher resolution layers for more precise segmentation. After a concatenation, each followed by a rectified linear unit, a dropout (Srivastava et al., 2014) layer with probability 0.5 is added for regularization to reduce overfitting. Finally, a sigmoid operation is applied to calculate probability for each class to obtain the output segmentation map  $\mathbf{O}$ . Fig.2 illustrates the segmentation network. More deconvolution layers were not added as they increased computation with no improvement in performance.

The same architecture is used for both the discriminator networks. It consists of five convolution layers with filter size  $4 \times 4$ , each followed by *LeakyRelu()*, Batch normalization layers (Ioffe and Szegedy, 2015). The first four convolution layers are followed by  $3 \times 3$  max pooling operations. Finally, a sigmoid layer is used to scale the output to probabilities i.e. in the range  $[0, 1]$ .

## 4. Experiments

### 4.1. Datasets

The DRIVE dataset (Staal et al., 2004) is used for training and testing for the retinal vessels which contains 40 images. The IDRiD dataset (Porwal et al., 2018) contains 81 retinal fundus images with annotations of the optic disk, microaneurysms, haemorrhages, hard exudates and soft exudates. Both datasets are already divided into two standard splits for training and testing. The IDRiD images of size  $4288 \times 2848$  px were downsampled to  $880 \times 584$  px while preserving the aspect ratio. DRIVE images were zero padded to a size of  $880 \times 584$  px.

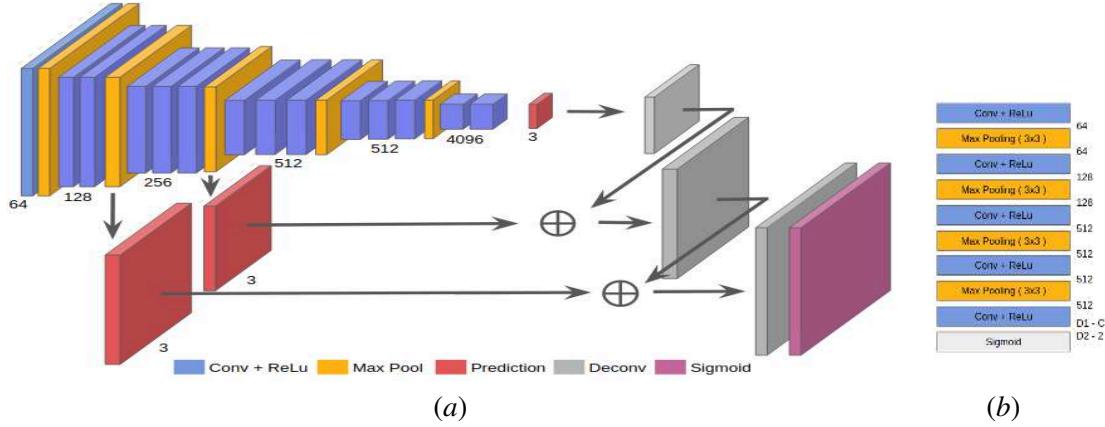


Figure 2: (a) The segmentation network architecture; (b) Basic architecture of both the discriminators: the last convolution layer has output of  $C - 1$  channels for Discriminator 1 (D1) and 2 channels for Discriminator 2 (D2)

#### 4.2. Training Setup

As data available for training a deep network is less, to overcome this limitation various augmentations are used to increase the number of retinal images for gaining better generalization and preventing overfitting. Images are randomly flipped by probability  $p$  generated using a uniform distribution or rotated by an angle  $\theta$  chosen from an uniform distribution  $[-180, 180]$  degrees.

The segmentation network uses the initial layers of VGG network (Simonyan and Zisserman, 2015) and those layers are initialized with the weights of a pretrained VGG model. As the dataset size is small, this initialization improves the results by a considerable margin. The vanilla FCN segmentation model is considered as a baseline. Experiments are conducted where both vanilla segmentation network and the adversarial approach are trained and compared resulting in proposed network achieving higher IOU score.

Adam optimizer (Kingma and Ba, 2014) with learning rate  $10^{-4}$  and  $\beta = 0.9$  is used for training all the three networks. The trade-off coefficient  $\lambda$  in (3) is set to 10. In every epoch the discriminators are trained first and then these losses summed with the segmentation loss is backpropagated through the segmentation network. Early stopping of the training based on the validation loss is adopted to prevent overfitting.

#### 4.3. Evaluation

Performance evaluation of segmentation is carried out by considering : sensitivity (SE), specificity (SP), accuracy (ACC), F-score (F1) and Area under the curve (AUC) for vessels, for comparison to previous methods. The AUC is calculated as the area under the SE versus  $(1-SP)$  plot. For lesions, the area under precision (PPV) vs sensitivity (SE) curve is computed, while for optic disk the Jaccard index (IOU) is measured so as to compare with the reported results of the Diabetic Retinopathy Segmentation challenge. The code for trying out inference has been released at <https://github.com/oindrilasaha/multitask-retinal-segmentation>.

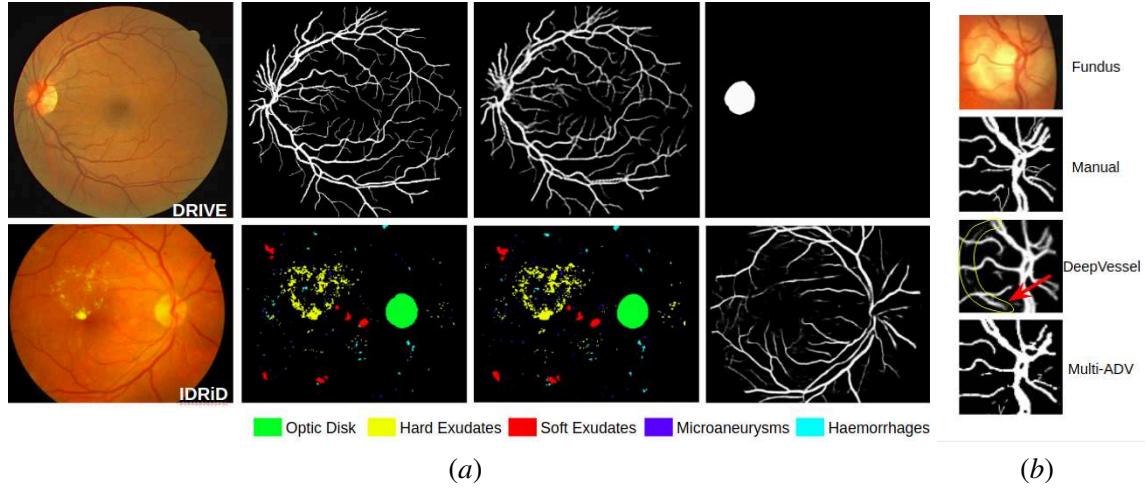


Figure 3: (a) Qualitative results: First column shows the retinal image, second column shows the manual annotation used as ground truth, third column shows the results obtained by the proposed method for the annotated classes in original dataset, and fourth column shows prediction of class not annotated but present in the image; (b) Comparison with DeepVessel (Fu et al., 2016)

## 5. Results

The proposed method is compared with the existing methods for vessel segmentation in Table 1. SE, SP, F1 and ACC are reported as the maximum value over thresholds. The FCN models trained without discriminator, show inferior performance with respect to Multi-ADV suggesting that the adversarial framework improves quality of segmentation. Our method outperforms existing vessel segmentation methods in literature. Quantitatively, for lesion and optic disc segmentation, the proposed method is compared to the best performing teams from the Diabetic Retinopathy challenge leaderboard<sup>3</sup> using the metrics used in the challenge for ranking in Table 2. Fig. 3a illustrates the qualitative results of the Multi-ADV model on DRIVE and IDRiD fundus images. It also shows how the method predicts classes which are not annotated for either dataset. It predicts optic disk for DRIVE images and retinal vessels for IDRiD images too. Fig. 3b compares the result of DeepVessel (Fu et al., 2016) to our method. DeepVessel incorrectly predicts the optic disk boundary as a retinal vessel while it was claimed that their method eliminates incorrect over-segmentation. Our method not only prevents over-segmentation but has also detected more finer vessels than DeepVessel.

## 6. Conclusion

In this paper, a multiadversary based fully convolutional neural network is proposed for retinal anatomy and pathology segmentation using weakly labelled fundus images. A FCN with skip connections is used as the segmentor network. The skip connections propagate context information to higher resolution layers for more precise segmentation. Two discriminators are used: the first discriminator contributes to making the segmentation result as realistic as possible while the second

3. <https://idrid.grand-challenge.org/Leaderboard/>

Table 1: Comparison with existing models for Vessel Segmentation

Method	SE	SP	ACC	F1	AUC
Azzopardi and Petkov (2013)	0.7655	0.9704	0.9442	-	0.9614
Liskowski and Krawiec (2016)	0.7811	0.9807	0.9535	-	0.9790
DeepVessel (Fu et al., 2016)	0.7603	-	0.9523	-	-
Orlando et al. (2017)	0.7897	0.9684	0.9454	0.7857	0.9506
Alom et al. (2018)	0.7792	0.9813	0.9556	-	0.9784
Maninis et al. (2016)	-	-	-	<b>0.8220</b>	-
FCN	0.7731	0.9724	0.9467	0.7612	0.9588
Multi-ADV	<b>0.7906</b>	<b>0.9839</b>	<b>0.9641</b>	0.7925	<b>0.9812</b>

Table 2: Comparison with IDRiD leaderboard for Optic Disk and Lesions

Class	Metric	Existing	Proposed	ACC - Proposed
Optic Disk	Jaccard Index (IOU)	0.9338	<b>0.9644</b>	0.9829
Microaneurysms	Area under PPV vs SE	0.5017	<b>0.5504</b>	0.9034
Haemorrhages	Area under PPV vs SE	0.6804	<b>0.7338</b>	0.9655
Soft Exudates	Area under PPV vs SE	0.6995	<b>0.7118</b>	0.9618
Hard Exudates	Area under PPV vs SE	<b>0.8850</b>	0.8698	0.9771

discriminator works to enforce segmentation of classes not annotated for a particular image. Results indicates that proposed approach outperforms other state-of-the-arts. The model also predicts classes that were originally not annotated in a given dataset by learning from other dataset.

## References

- M Usman Akram, Shehzad Khalid, Anam Tariq, Shoab A Khan, and Farooque Azam. Detection and classification of retinal lesions for grading of diabetic retinopathy. *Comp. Biol., Med.*, 45: 161–171, 2014.
- Bashir Al-Diri, Andrew Hunter, and David Steel. An active contour model for segmenting and measuring retinal vessels. *IEEE Trans. Med. imaging*, 28(9):1488–1497, 2009.
- Sharib Ali, Désiré Sidibé, Kedir M Adal, Luca Giancardo, Edward Chaum, Thomas P Karnowski, and Fabrice Mériaudeau. Statistical atlas based exudate segmentation. *Comp. Med. Imaging, Graph.*, 37(5-6):358–368, 2013.
- Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*, 2018.
- George Azzopardi and Nicolai Petkov. Automatic detection of vascular bifurcations in segmented retinal images using trainable cosfire filters. *Pattern Recognition Letters*, 34(8):922–933, 2013.

- Brad Bowling. *Kanski's Clinical Ophthalmology E-Book: A Systematic Approach*. Elsevier Health Sciences, 2015.
- Eid Emary, Hossam M Zawbaa, Aboul Ella Hassani, Gerald Schaefer, and Ahmad Taher Azar. Retinal vessel segmentation based on possibilistic fuzzy c-means clustering optimised with cuckoo search. In *Proc. Int. Jt. Conf. Neur. Networks*, pages 1792–1796, 2014.
- Huazhu Fu, Yanwu Xu, Stephen Lin, Damon Wing Kee Wong, and Jiang Liu. Deepvessel: Retinal vessel segmentation via deep learning and conditional random field. In *Proc. Int. Conf. Med. Image Comput., Comp.-Assist. Interv.*, pages 132–139, 2016.
- Mahendran Gandhi and R Dhanasekaran. Diagnosis of diabetic retinopathy using morphological process and svm classifier. In *Proc. Int. COnf. Comm., Signal Process.*, pages 873–877, 2013.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, 2016.
- María García, Clara I Sánchez, María I López, Daniel Abásolo, and Roberto Hornero. Neural network based detection of hard exudates in retinal images. *Comp. Methods, Programs, Biomed.*, 93(1):9–19, 2009.
- A Haddouche, Mouloud Adel, Monique Rasigni, J Conrath, and Salah Bourennane. Detection of the foveal avascular zone on retinal angiograms using markov random fields. *Digital Signal Processing*, 20(1):149–154, 2010.
- Tsuyoshi Inoue, Yuji Hatanaka, Susumu Okumura, Chisako Muramatsu, and Hiroshi Fujita. Automated microaneurysm detection method based on eigenvalue analysis using hessian matrix in retinal fundus images. In *Proc. An. Conf. IEEE Engg., Med., Biol. Soc.*, pages 5873–5876, 2013.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. Int. Conf. Machine Learn.*, pages 448–456, 2015.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- Paweł Liskowski and Krzysztof Krawiec. Segmenting retinal blood vessels with deep neural networks. *IEEE Trans. Medi. Imaging*, 35(11):2369–2380, 2016.
- Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE/CVF Conf. Comp. Vis., Patt. Recog.*, pages 3431–3440, 2015.
- Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.
- Carmen Alina Lupascu, Domenico Tegolo, and Emanuele Trucco. Fabc: retinal vessel segmentation using adaboost. *IEEE Trans. Inf. Tech. Biomed.*, 14(5):1267–1274, 2010.
- Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Deep retinal image understanding. In *Proc. Int. Conf. Med. Image Comput., Comp.-Assist. Interv.*, pages 140–148, 2016.

- Muthu Rama Krishnan Mookiah, U. Rajendra Acharya, Chua Kuang Chua, Choo Min Lim, E. Y. K. Ng, and Augustinus Laude. Computer-aided diagnosis of diabetic retinopathy: A review. *Comput. Biol. Med.*, 43(12):2136–2155, Dec. 2013.
- Martin M Nentwich and Michael W Ulbig. Diabetic retinopathy-ocular complications of diabetes mellitus. *World J. Diabetes*, 6(3):489, 2015.
- Jan Odstrcilik, Radim Kolar, Attila Budai, Joachim Hornegger, Jiří Jan, Jiri Gazarek, Tomas Kubena, Pavel Cernosek, Ondrej Svoboda, and Elli Angelopoulou. Retinal vessel segmentation by improved matched filtering: Evaluation on a new high-resolution fundus image database. *IET Image Process.*, 7:373–383, 06 2013.
- Jose Orlando, Elena Prokofyeva, and Matthew Blaschko. A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. *IEEE Trans. Biomed. Engg.*, 64(1):16–27, Jan 2017.
- Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian Diabetic Retinopathy Image Dataset (IDRID). IEEE Dataport, 2018.
- Elisa Ricci and Renzo Perfetti. Retinal blood vessel segmentation using line operators and support vector classification. *IEEE Trans. Medi. Imaging*, 26(10):1357–1365, 2007.
- Ana G Salazar-Gonzalez, Djibril Kaba, Yongmin Li, and Xiaohui Liu. Segmentation of the blood vessels and optic disk in retinal images. *IEEE J. Biomed. Health Inf.*, 18(6):1874–1886, 2014.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learn. Rep.*, 2015.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.
- Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging*, 23 (4):501–509, 2004.
- Jen Hong Tan, Hamido Fujita, Sobha Sivaprasad, Sulatha V Bhandary, A Krishna Rao, Kuang Chua Chua, and U Rajendra Acharya. Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network. *Information Sciences*, 420:66–76, 2017.
- Thomas Teng, Martin Lefley, and D Claremont. Progress towards automated diabetic ocular screening: a review of image analysis and intelligent systems for diabetic retinopathy. *Med., Biol., Engg., Comp.*, 40(1):2–13, 2002.
- Li Wang, Abhir Bhalerao, and Roland Wilson. Analysis of retinal vasculature using a multiresolution hermite model. *IEEE Trans. Med. Imaging*, 26(2):137–152, 2007.
- Bob Zhang, Lin Zhang, Lei Zhang, and Fakhri Karray. Retinal vessel extraction by matched filter with first-order derivative of gaussian. *Comp. Biol., Med.*, 40(4):438–445, 2010.

## Appendix

A few experimental results are demonstrated on other datasets. Inference only was performed on these images using our existing model without re-training it with any images from these datasets.

### Inference on Vessel dataset

HRF dataset <sup>4</sup> was used for inference for vessels class. The results are presented in Table 3. It is worthwhile to note that despite of the network not having had seen HRF images during training, it performs almost equivocal with slightly higher sensitivity and accuracy as compared to the competitive prior art for the task of vessel segmentation.

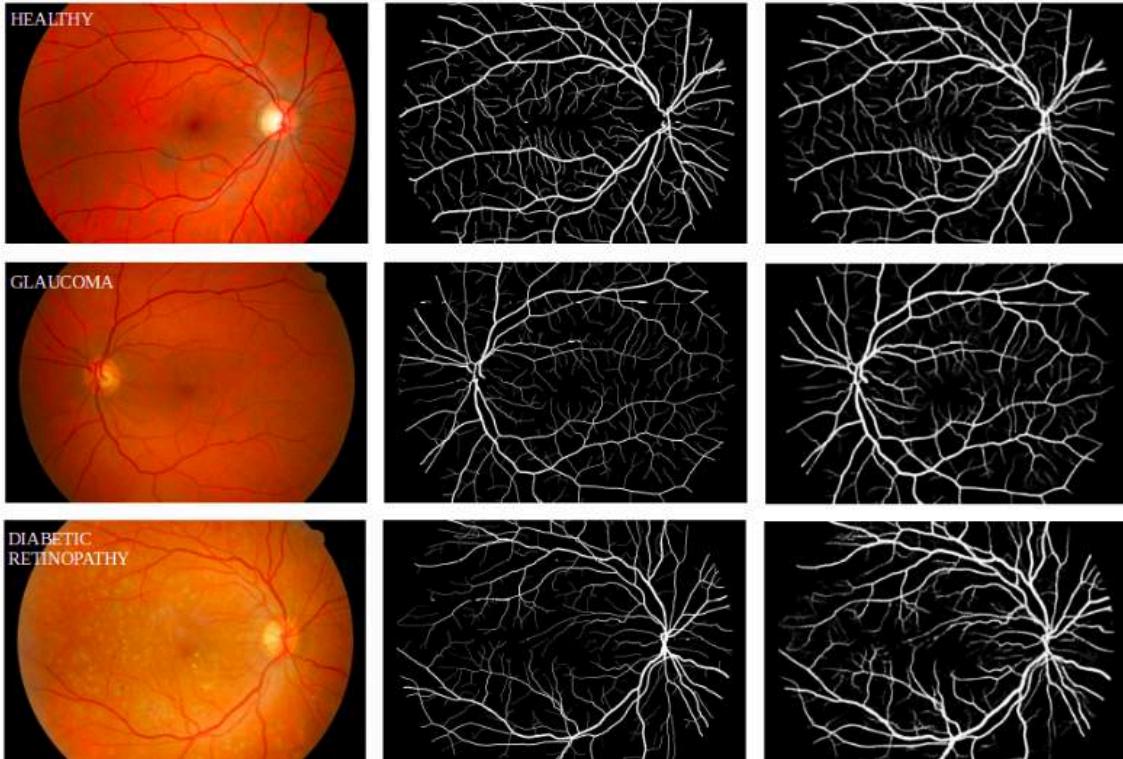


Figure 4: Vessel Segmentation on HRF dataset. Left: Fundus Image, center: Ground Truth, right: Prediction

---

4. <https://www5.cs.fau.de/research/dat/fundus-images/>

Table 3: Performance evaluation on HRF dataset

Method	SE	SP	ACC	F1
Proposed	0.7891	0.9642	0.9610	0.7552
Odstrcilik et al. (2013)	0.7741	0.9669	0.9493	-

### Inference on pathology datasets

Similar to the approach in for vessel segmentation above, only inferencing is employed on e-optha<sup>5</sup> for exudates and microaneurysms (Table 4), and REFUGE<sup>6</sup> for optic disk (Table 5). For the REFUGE dataset, the results were compared with the challenge leaderboard<sup>7</sup> results.

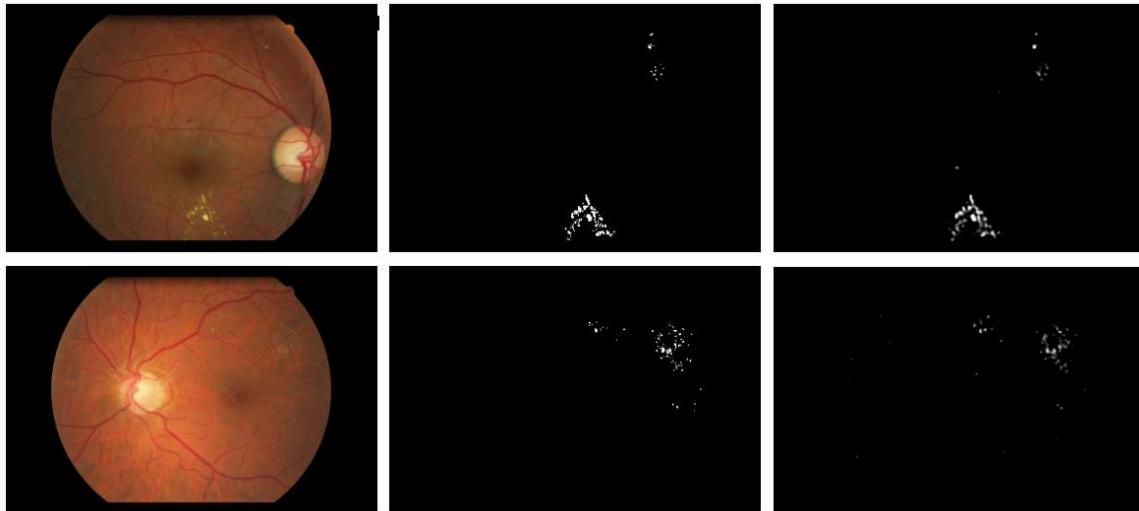


Figure 5: Hard Exudates Segmentation on eoptha-EX dataset. Left: Fundus Image, center: Ground Truth, right: Prediction

Table 4: Performance evaluation on eoptha dataset

Dataset	Area under Precision Recall
eoptha-EX	0.8235
eoptha-MA	0.2500

5. <http://www.adcis.net/es/Descargar-Software-Base-De-Datos-Terceros/E-Ophtha.html>

6. <https://refuge.grand-challenge.org>

7. <https://refuge.grand-challenge.org/leaderboard>

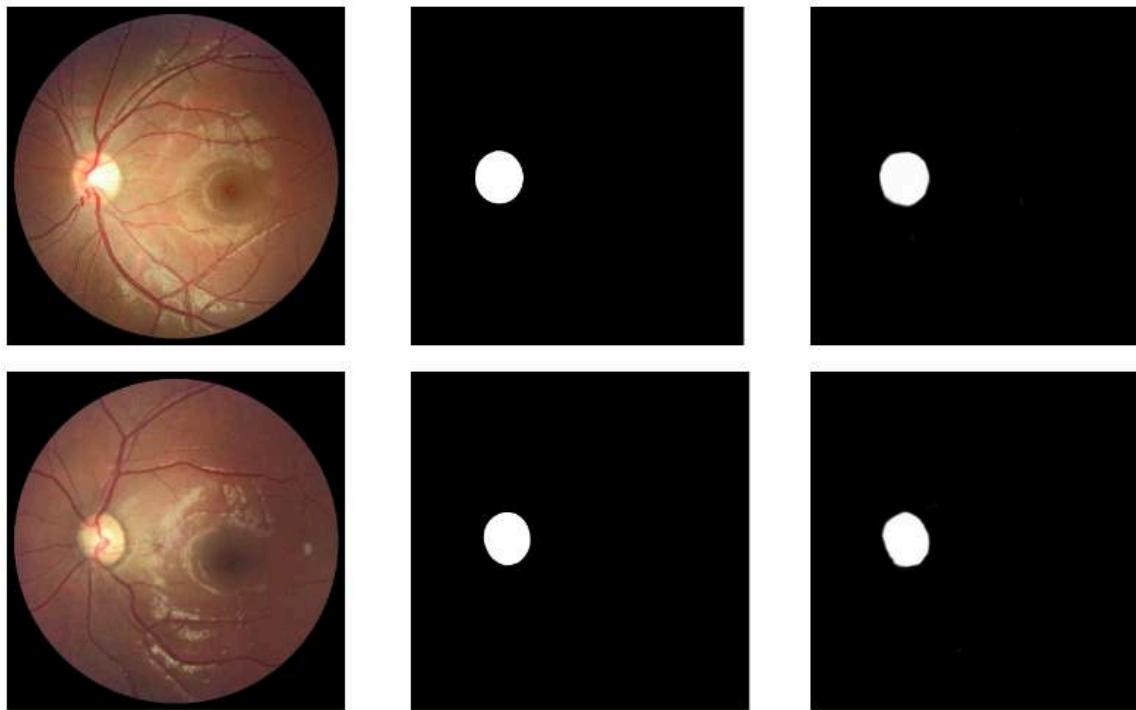


Figure 6: Optic Disk Segmentation on REFUGE dataset. Left: Fundus Image, center: Ground Truth, right: Prediction

Table 5: Performance evaluation on REFUGE dataset

Method	F1	Jaccard
Proposed	0.9286	0.9202
REFUGE leaderboard	0.958	-

# MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty

**Richard Shaw<sup>1,2</sup>**

**Carole Sudre<sup>2,1,3</sup>**

**Sebastien Ourselin<sup>2</sup>**

**M. Jorge Cardoso<sup>2</sup>**

RICHARD.SHAW.17@UCL.AC.UK

CAROLE.SUDRE@KCL.AC.UK

SEBASTIEN.OURSELIN@KCL.AC.UK

M.JORGE.CARDOSO@KCL.AC.UK

<sup>1</sup> Department of Medical Physics and Biomedical Engineering, University College London, UK

<sup>2</sup> School of Biomedical Engineering and Imaging Sciences, King’s College London, UK

<sup>3</sup> Dementia Research Centre, Institute of Neurology, UCL, UK

## Abstract

Patient movement during the acquisition of magnetic resonance images (MRI) can cause unwanted image artefacts. These artefacts may affect the quality of diagnosis by clinicians and cause errors in automated image analysis. In this work, we present a method for generating realistic motion artefacts from artefact-free data to be used in deep learning frameworks to increase training appearance variability and ultimately make machine learning algorithms such as convolutional neural networks (CNNs) robust to the presence of motion artefacts. We model patient movement as a sequence of randomly-generated, ‘de-means’, rigid 3D affine transforms which, by resampling artefact-free volumes, are then combined in k-space to generate realistic motion artefacts. We show that by augmenting the training of semantic segmentation CNNs with artefacted data, we can train models that generalise better and perform more reliably in the presence of artefacted data, with negligible cost to their performance on artefact-free data. We show that the performance of models trained using artefacted data on segmentation tasks on real-world test-retest image pairs is more robust. Finally, we demonstrate that measures of uncertainty obtained from motion augmented models reflect the presence of artefacts and can thus provide relevant information to ensure the safe usage of deep learning extracted biomarkers in clinics.

## 1. Introduction

Patient movement during the acquisition of magnetic resonance images (MRI) can result in unwanted image artefacts, which manifest as blurring, ringing or ghosting effects, depending on both timing and spatial changes during a scan (Wood and Henkelman, 1985). Motion artefacts can affect the interpretability of images, potentially affecting the quality of diagnosis by clinicians and/or leading to increased cost if the images are judged unusable and the acquisition has to be repeated. Artefacts can also affect the performance of post-processing algorithms, and it has been shown that motion artefacts consistently affect segmentation measurements on structural MR images. Additionally, in the context of research cohorts, artefacts may lead to inclusion bias in statistical analysis as more impaired subjects tend to have difficulties staying still, resulting in poorer quality scans more likely to be excluded (Wylie et al., 2014). Even if included, biomarker measures may be biased by artefacts leading to spurious findings (Alexander-Bloch et al., 2016).

The type of motion artefacts that appear in MR images depends on the amount and timing of patient movement with regards to the k-space scan trajectory. Movements at the k-space centre

correspond to low image frequencies and result in ghosting artefacts, where the image is repeated, as does quasi-periodic motion e.g. respiration (Zaitsev et al., 2015). Movements toward the edges of the k-space corresponding to the acquisition of high image frequencies, often result in ringing artefacts. Most commonly observed MR motion artefacts result in minor blurring due to small movements spanning a range of frequencies during k-space acquisition. Additionally, motion artefact appearance depends on the k-space scanning strategy and notably whether the acquisition is performed in 2D or 3D.

Prior work on motion artefacts in MRI has mostly attempted to design ways of correcting for them (Usman et al., 2013). This work addresses the problem of motion artefacts under a different perspective – attempting to make automated systems of image analysis more robust to their presence. In recent years, deep learning frameworks have demonstrated high performance when applied to segmentation and classification tasks. In a deep learning setup, data augmentation is a classical way to artificially increase data variability and thus increase the networks potential for generalisation (Çiçek et al., 2016). While classical data augmentation usually involves random geometric transformations and/or intensity changes, biologically and physically plausible augmentation models would be beneficial.

Some prior work has been done in this domain. Meding et al. (Meding et al., 2017) used convolutional neural networks to classify MR magnitude images as artefacted or not. Going beyond the binary classification task, Duffy et al. (Duffy et al., 2018), also using CNNs, attempt to learn how to retrospectively remove artefacts from MR images. Their network, trained on synthetic data proposes an unrealistic motion model that is limited to axial translation. Using a Generative Adversarial Network, Armanious et al. propose MedGAN (Armanious et al., 2018) with the objective of “translating” motion-corrupted MR images to their corresponding motion-free images, but restrict their work to 2D slices. Pawar et al. (Pawar et al., 2018), with the objective of learning to remove artefacts, model 3D motion in the image domain and reconstruct the k-space from multiple resampled images using however only 2D axial slices. In contrast to these approaches, we argue that it is ultimately more useful to optimise frameworks in an end-to-end manner, rather than generating intermediate clean images, thus enforcing robustness to artefacts at the level of the internal representation of the data. Such strategy inherently avoids caveats of GANs, that may wrongly introduce non-existing information (hallucination), or of artefact removal strategies that may only account for part of the present artefacts thus resulting in data that is unusable for further processing. Additionally, end-to-end learning allows for model artefact-induced task uncertainty to be learned directly from raw artefacted inputs.

## 2. Motion Artefact Model

Our proposed method for generating motion artefacts is illustrated in Figure 1. The procedure is summarised by the following steps: (1) Generate a random movement model by sampling different probability distribution functions. (2) De-mean the generated movement transforms. (3) Resample the artefact-free volume according to the de-meanned movement model. (4) Reconstruct a composite k-space from the k-spaces of multiple resampled volumes. (5) Transform the k-space to the image domain to produce final artefacted sample.

Taking each step in turn, we first sample movements from PDFs, modelling head motion as a sequence of independent small and large motions (e.g. twitches/nodding). We sample  $N$  movements from a Poisson distribution - small movements are assumed to occur more often and large move-

ments less frequently - and the time  $t$  in k-space at which each movement occurs from a uniform distribution (assuming k-space scans in the phase encoding direction). Modelling each movement as a 3D affine  $A$  matrix comprising a rigid 3D rotation and translation in the image domain, the rotation is sampled from between  $(-30^\circ, 30^\circ)$  and the translation between  $(-10\text{mm}, 10\text{mm})$  in all three axes. The sequence of movement transforms  $\{A\}_{i=0}^N$  is combined incrementally in log-Euclidean space (Alexa, 2002), allowing for the linear combination of transforms. To apply affine matrix  $A$  in log-Euclidean space, we use the matrix exponential  $\exp_M(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}$ , and corresponding matrix logarithm.

With the motion model defined, the second step is to ‘de-mean’ the movements. When applying our augmentation model to a clean MR volume  $I_0$ , we expect the barycenter of the imaged object to remain in approximately the same position within the 3D volume. This is achieved by ‘de-meaning’ each affine transform  $A_i$  by pre-multiplying by the inverse of the average transform  $A_{avg}$ , computed as the weighted sum of the sequence  $N$  affine transformations in log-Euclidean space, given by Equation (1),

$$A_{avg} = \exp_M\left(\sum_{i=0}^N \hat{w}_i \log_M(A_i)\right), \quad (1)$$

where  $\hat{w}_i$  is a weighting given to the  $i$ -th movement. Since movements at different parts of the k-space contribute different spatial frequencies, we weight each  $A_i$  by its signal contribution to the final image. This means that movements at the k-space centre (low frequencies) have a higher weight since their impact on the final 3D position of the brain, and the overall Fourier power spectrum, is much greater. Each weight is estimated by masking the 3D k-space of  $I_0$  with a binary mask  $M_i$  corresponding to the k-space elements of the  $i$ -th movement, transforming back to the image domain, and summing the resulting voxel intensities, as given by  $w_i = \sum_{\text{voxels}} \mathcal{F}^{-1}(M_i \odot \mathcal{F}(I_0))$ , with the weights then normalised to sum to 1.

The third step is to apply each ‘de-means’ affine transform to the original artefact-free image volume  $I_0$  and resample using b-spline interpolation. We always resample the original image to reduce propagating interpolation errors and we apply edge-padding to mitigate edge effects. Fol-

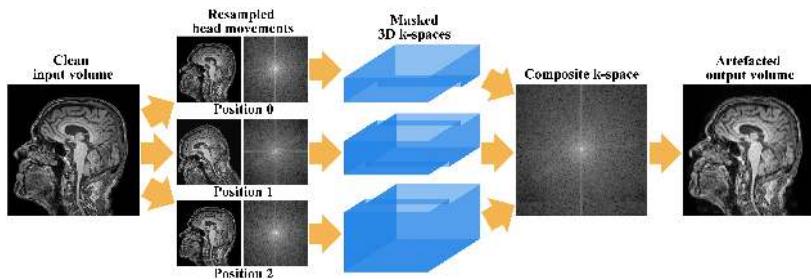


Figure 1: Motion artefact augmentation framework: The artefact-free volume is resampled according to a randomly sampled movement model, defined by a sequence of ‘de-means’ 3D affine transforms. Their 3D Fourier transforms are combined to form a composite k-space, which is transformed back to the image domain producing the final artefacted volume.

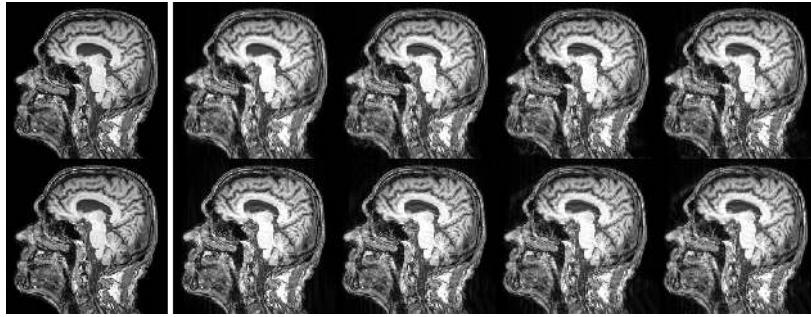


Figure 2: Simulated motion artefacts due to nodding motion. Top row: artefacts due to increasing amounts of movement from left to right. Bottom row: artefacts due to changing the time at which the movement occurs, earlier in the k-space scan trajectory from left to right, therefore retaining higher image frequencies. Best viewed zoomed in on digital copy.

lowing each transformation, we compute the 3D Fourier Transform of each resampled image. The fourth step is to combine the 3D Fourier transforms corresponding to each position of the brain in the sequence, joined together at sampled times  $t$ , forming a complete k-space of the scan containing movement. Finally, the inverse 3D Fourier Transform of the composite k-space is derived, and the magnitude image provides the final artefact sample. Examples of our artefact augmentation are shown in Figure 2.

### 3. Experiments

We evaluate our motion artefact augmentation model on both simulated and real-world data containing artefacts in the context of three segmentation tasks: cortical gray matter (CGM), hippocampus and total intracranial volume (TIV).

#### 3.1. Network architecture and implementation details

We used the HighResNet (Li et al., 2017) architecture implemented in NiftyNet (Gibson et al., 2018), with Dice loss (Sudre et al., 2017), patch size of  $80^3$  and batch size 1, trained on a single GPU with Adam optimiser (Kingma and Ba, 2014) and a learning rate of  $10^{-4}$ . In the context of segmentation, due to imbalance between foreground and background elements, the sampling strategy is essential to training performance, so we use weighted patch sampling with higher weight at regions defined by the blurred ground-truth label, such that the foreground/background weight ratio is roughly equal to the ratio of foreground/background voxels. Each model was trained until overfitting was observed or reaching 100,000 iterations.

#### 3.2. Simulated Dataset

For experiments on simulated data, we use 272 MPRAGE scans from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) and generate 15 artefact volumes per scan. The data was split into

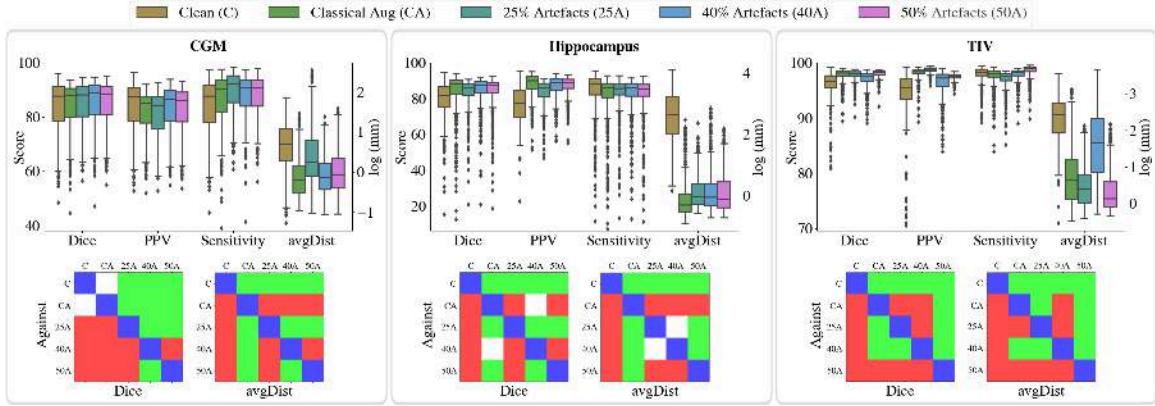


Figure 3: Segmentation results on simulated data for CGM, hippocampus and TIV across models. Top: Boxplots of different error metrics for 5 models with different augmentation: clean, classically augmented, 25%, 40%, 50% artefacts. Bottom: Bonferroni corrected pairwise Wilcoxon tests for Dice and average distance between column and row models - Green: significantly better, White: no statistically significant difference; Red: significantly worse.

80% training, 10% validation and 10% testing and CNNs were trained to segment CGM, hippocampus and TIV. For each segmentation task, five models were trained with varying levels and types of augmentation. One was trained only on ‘clean’ data, i.e. the original artefact-free scans. Another was trained with ‘classical’ augmentation, consisting of random rotation, translation and scaling. The remaining three models were trained with increasing amounts of motion artefact augmentation: where 25%, 40% and 50% of images seen in the training set contain artefacts, in addition to classical augmentations. Each model includes bias field augmentation by default to account for variability in image intensity across samples. All models are tested on the same hold out test set containing both clean and artificially artefacted data. Segmentation performance for these three tasks across all models is evaluated with Dice score, positive predictive value, sensitivity and average distance metrics and presented in the first row of Figure 3. Results of Bonferroni corrected matched pair Wilcoxon tests between models are presented on the bottom row.

### 3.3. Real-world Dataset

Robustness of CNNs trained with the proposed motion augmentation to real-world motion artefacts was then evaluated in a test-retest setting. 106 quality-controlled pairs of MPRAGE test-retest images from the ADNI dataset on which only one of the images was considered artefacted were used for this purpose. Image pairs were rigidly registered together in a groupwise space to avoid interpolation bias. For comparison purposes, a benchmark label fusion algorithm was used to perform the segmentation tasks on each pair of images (Cardoso et al., 2015). For each trained model and the benchmark method, Dice score, positive predictive value, sensitivity and average distance were used as evaluation measure between test and retest images, with the results obtained on the clean image being used as reference. Figure 4 presents in the top row the corresponding boxplots for each

segmentation task, while the second row displays the Bonferroni corrected matched-pair Wilcoxon tests across models.

#### 4. Task-specific Uncertainty Estimation

Deep learning models for segmentation tasks classically provide for each voxel a point-estimate probability of belonging to a certain class. Being able to provide in addition a calibrated measure of the uncertainty of a given prediction has become essential in applications for which safety is paramount such as medical applications.

As theorised by Gal and Ghahramani (Gal and Ghahramani, 2016), uncertainty can be estimated by sampling at inference time from multiple outputs of the network trained with dropout. Adapting the approach from (Eaton-Rosen et al., 2018), uncertainty over the segmentation result is obtained as the variance over the predictions made from multiple forward passes of the dropout network. For training, the dropout rate was set at 0.5 everywhere except the initial layer, which was set to 0.05. Mean and variance results obtained on the CGM segmentation task for the aforementioned models trained with dropout are shown in Figure 5. In models trained with motion augmentation, higher variance of predictions are observed in artefactected regions, especially close to the cortical surface.

To further investigate the behaviour of segmentation uncertainty estimation in the presence of motion artefacts, with respect to the type of augmentation applied at training, per-voxel Kullback-Leibler divergence (KLD) maps comparing the sampled distributions for clean and artefactected images were calculated, as shown in the bottom row of Figure 5. By associating KLD with uncertainty, as measured by the sampled variance (std), we can examine this relationship for each model and mode of augmentation, visualised by the histogram plots of uncertainty on the artefactected image vs KLD in Figure 6.

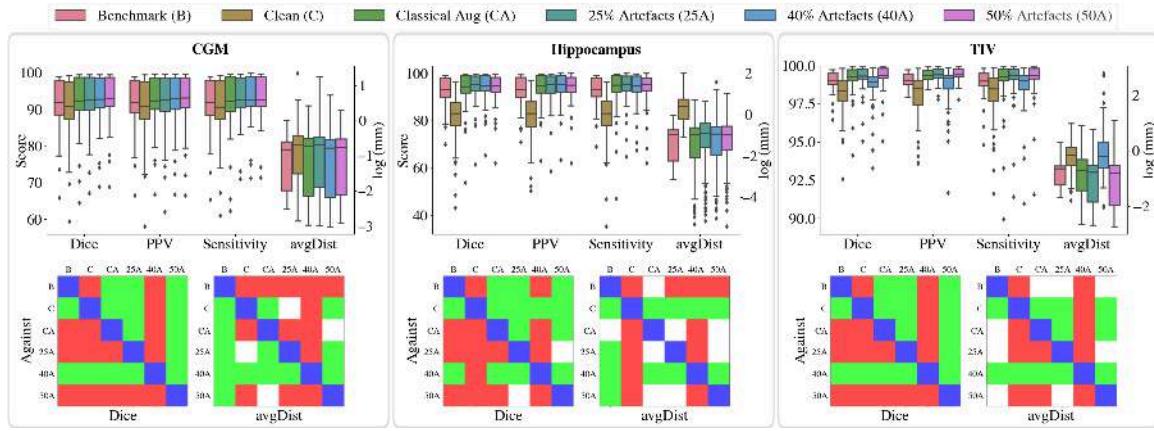


Figure 4: Segmentation robustness on real-world test-retest data for CGM, hippocampus and TIV across models. Top: For each task, boxplots of different error metrics for the 5 models in addition to a benchmark method. Bottom: Bonferroni corrected pairwise Wilcoxon tests for Dice and average distance between column and row models - Green: significantly better, White: no statistically significant difference; Red: significantly worse.

Different modes of association between uncertainty and KLD can be interpreted as follows: 1) low std - low KLD: the model gives similar predictions on clean and artefacted images in a confident fashion; 2) high std - low KLD: the model provides highly similar distributions but overestimates uncertainty 3) low std - high KLD: the model provides mismatching answers with high confidence, a clinically unsafe behaviour 4) high std - high KLD: the probability distributions are different from each other but the model is aware that it cannot ascertain the results with certainty. Note that, in the

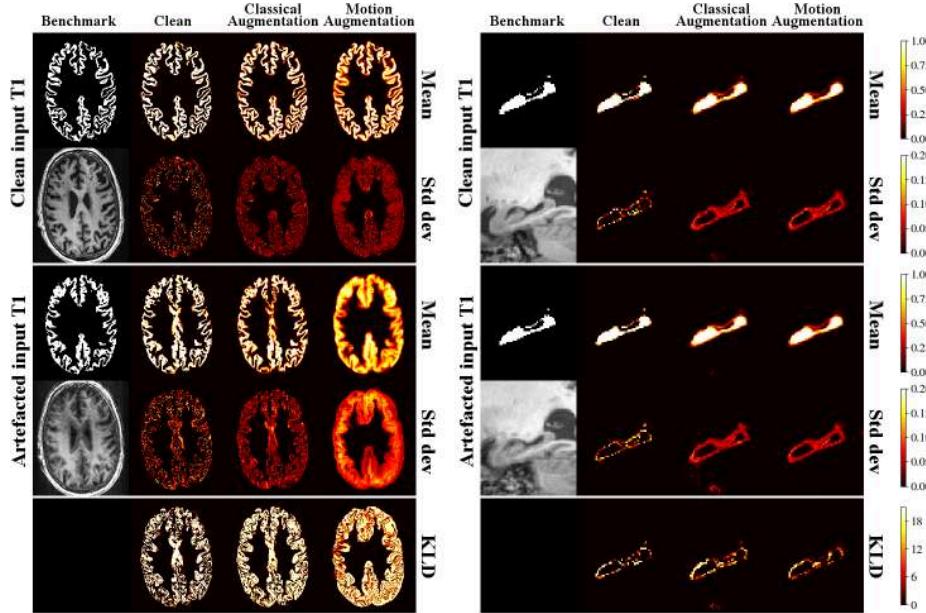


Figure 5: Per-voxel mean and uncertainty estimations on CGM and hippocampus segmentation tasks for clean (no augmentation), classically augmented and motion augmented models for a test-retest pair for which one scan is heavily artefacted. The segmentation produced by a benchmark method is shown for reference. Bottom row: KL-divergence (KLD) between the probability distributions produced by each model on clean and artefacted scans.

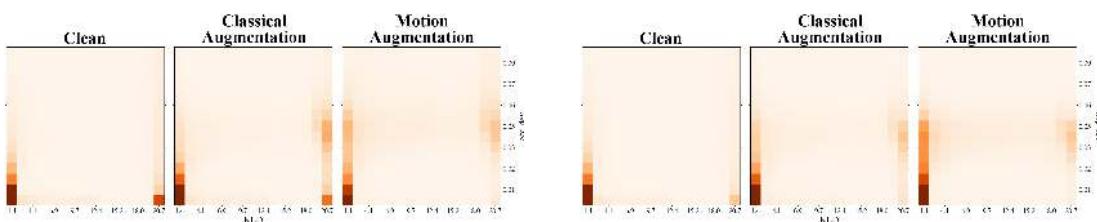


Figure 6: Histograms of per-voxel KLD associated to the uncertainty estimates as measured by the sampled variance, shown for models trained with different augmentations on (right) heavily artefacted and (left) lightly artefacted data.

presence of heavily artefacted images (Figure 6 left), models trained on clean (artefact-free) data or with only classical augmentation behave unsafely more often, i.e. more predictions with high KLD and low uncertainty. Models trained with motion augmentation were found to be safer.

## 5. Discussion and Conclusion

Considering the results on data with synthetic artefacts, in the tasks of CGM, hippocampus and TIV segmentation, models trained with motion artefact augmentation perform generally better than models without any augmentation or with only classical augmentation (rotation, translation and scaling). For CGM and TIV segmentation, in terms of Dice, PPV and sensitivity metrics, the model that observes the most artefacts during training (50% artefacts) consistently performs significantly better than the others, with a lower result variance. For the hippocampus, the benefit of artefact augmentation is less clear. This is likely related to the location of the object (medial brain), thus being less affected by extra-cranial fat ringing artefacts. For example, ringing artefacts mainly impact the cortical surface, and ghosting typically affects the TIV. For the average distance metric, the classically augmented model appears to perform better on CGM and hippocampus, whereas for TIV the motion artefact model is statistically significantly better.

On real-world data we observe a similar benefit to performance when training with simulated data. In terms of Dice score coefficient, PPV and sensitivity, the motion augmented models mostly perform better. This suggests the proposed motion artefact generation is realistic and contributes to increased robustness to artefacts of models trained with this augmentation. Additionally, it appears that the larger the artefactual variability encountered at training the better the performance of the model. Although artefact augmentation shows promising results for segmentation, there are limitations with the proposed model:

First, the augmentation model assumes a valid segmentation exists, but this may not always be true. With heavily artefacted scans caused by extreme movements, it is difficult to say with certainty where the true segmentation should be. If the subject's head was in one place for 50% of the scan and in another position for the remaining time, where should the ground-truth segmentation be located? In this case, an uncertain segmentation is the only hypothetically correct answer.

Second, our CNN models are parameter-deprived due to memory constraints, as training with artefacts decreases inference performance on clean data. Note, however, that this drop in performance on clean data is not statistically significant, while providing significant improvements on artefacted data. While performance is a key goal, robustness to data artefacts is paramount to enable the safe clinical translation of such technique.

Third, our motion model is randomly sampled from PDFs, but human motion in MRI is not completely random and certain motions are more common, e.g. nodding when the patient swallows. Therefore our simulated dataset is not entirely representative of the distribution of observed motion artefacts. With more consideration of the types of movements that occur, adaptation of the model could see potential increased performance on real data.

Notwithstanding these observations, our main contributions are threefold: Firstly, a realistic, fully 3D, motion model of MRI acquisitions to augment training data, improving the performance and robustness of semantic segmentation CNNs to real-world artefacts. Training on simulated artefacts has been shown to successfully translate to improved performance on real-world artefacts, while the performance on artefact-free data is largely unaffected by the use of augmented data during training. Secondly, by training the different tasks end-to-end with motion augmentation, a

new internal data representation is created allowing the model to become robust to the presence of noise, instead of requiring an explicit intermediate step of artefact removal likely to destroy important image information. Lastly, our augmentation model provides more calibrated and informative uncertainty estimates for segmentation predictions in the presence of real-world motion-corrupted data. This is of utmost importance when addressing the question of safe clinical translation of such models.

What humans deem acceptable scan quality for radiological assessment is different to the quality required for automated analysis. With this in mind, we observe that scan quality is intrinsically related to the task being solved. This observation, as opposed to a human-perceived notion of image-wide scan quality, is a concept rarely recognised by machine learning researchers, systems and datasets.

## References

- Marc Alexa. Linear combination of transformations. In *SIGGRAPH*, 2002.
- Aaron Alexander-Bloch, Liv S. Clasen, Michael Stockman, Lissa Ronan, Francois M. Lalonde, Jay N. Giedd, and Armin Raznahan. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo mri. *Human brain mapping*, 37(7):2385–97, 2016.
- Karim Armanious, Chenming Yang, Marc Fischer, Thomas Küstner, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *CoRR*, 2018.
- M. Jorge Cardoso, Marc Modat, Robin Wolz, Andrew Melbourne, David M. Cash, Daniel Rueckert, and Sébastien Ourselin. Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion. *IEEE Transactions on Medical Imaging*, 34:1976–1988, 2015.
- Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *MICCAI*, 2016.
- Ben A. Duffy, Wenlu Zhang, Haoteng Tang, Lu Zhao, Meng Law, Arthur W. Toga, and Hosung Kim. Retrospective correction of motion artifact affected structural mri images using deep learning of simulated motion. *MDL*, 2018.
- Zach Eaton-Rosen, Felix J. S. Bragman, Sotirios Bisdas, Sébastien Ourselin, and M. Jorge Cardoso. Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions. In *MICCAI*, 2018.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, volume 48 of *ICML*, pages 1050–1059, 2016.
- Eli Gibson, Wenqi Li, Carole H. Sudre, Lucas Fidon, Dzhoshkun I. Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, Tom Whyntie, Parashkev Nachev, Marc Modat, Dean C. Barratt, Sébastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. Niftynet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, 158:113–122, 2018.

- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- Wenqi Li, Guotai Wang, Lucas Fidon, Sébastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. On the compactness, efficiency, and representation of 3d convolutional networks: Brain parcellation as a pretext task. In *IPMI*, 2017.
- Kristof Meding, Alexander Loktyushin, and Michael Hirsch. Automatic detection of motion artifacts in mr images using cnns. In *42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, pages 811–815, 2017.
- Kamlesh Pawar, Zhaolin Chen, N Jon Shah, and Gary Egan. Moconet: Motion correction in 3d mprage images using a convolutional neural network approach. *arXiv e-prints*, art. arXiv:1807.10831, 2018.
- Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *DLMIA/ML-CDS@MICCAI*, 2017.
- Muhammad Usman, David Atkinson, Freddy Odille, Christoph Kolbitsch, Ghislain Vaillant, Tobias Schaeffter, Philip G. Batchelor, and Claudia Prieto. Motion corrected compressed sensing for free-breathing dynamic cardiac mri. *Magnetic resonance in medicine*, 70(2):504–16, 2013.
- Michael L. Wood and Mark Henkelman. Mr image artifacts from periodic motion. *Medical Physics*, 12(2):143–151, 1985.
- Glenn R. Wylie, Helen M. Genova, John DeLuca, Nancy D. Chiaravalloti, and James F. Sumowski. Functional magnetic resonance imaging movers and shakers: does subject-movement cause sampling bias? *Human brain mapping*, 35(1):1–13, 2014.
- Maxim Zaitsev, J. Piers Maclarens, and Michael F. Herbst. Motion artifacts in mri: A complex problem with many partial solutions. *Journal of magnetic resonance imaging : JMRI*, 42(4):887–901, 2015.

# A Hybrid, Dual Domain, Cascade of Convolutional Neural Networks for Magnetic Resonance Image Reconstruction

**Roberto Souza**<sup>1,2</sup>

ROBERTO.MEDEIROSDESO@UCALGARY.CA

**R. Marc Lebel**<sup>3</sup>

MARC.LEBEL@GE.COM

**Richard Frayne**<sup>1,2</sup>

RFRAYNE@UCALGARY.CA

<sup>1</sup>*Department of Radiology and Clinical Neuroscience, Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada*

<sup>2</sup>*Seaman Family MR Research Centre, Foothills Medical Centre, Calgary, AB, Canada*

<sup>3</sup>*General Electric Healthcare, Calgary, AB, Canada*

## Abstract

Deep-learning-based magnetic resonance (MR) imaging reconstruction techniques have the potential to accelerate MR image acquisition by reconstructing in real-time clinical quality images from k-spaces sampled at rates lower than specified by the Nyquist-Shannon sampling theorem, which is known as compressed sensing. In the past few years, several deep learning network architectures have been proposed for MR compressed sensing reconstruction. After examining the successful elements in these network architectures, we propose a hybrid frequency-/image-domain cascade of convolutional neural networks intercalated with data consistency layers that is trained end-to-end for compressed sensing reconstruction of MR images. We compare our method with five recently published deep learning-based methods using MR raw data. Our results indicate that our architecture improvements were statistically significant (Wilcoxon signed-rank test,  $p < 0.05$ ). Visual assessment of the images reconstructed confirm that our method outputs images similar to the fully sampled reconstruction reference.

**Keywords:** Magnetic resonance imaging, image reconstruction, compressed sensing, deep learning.

## 1. Introduction

Magnetic resonance (MR) is a non-ionizing imaging modality that possess far superior soft-tissue contrast compared to other imaging modalities (Nishimura, 1996). It allows us to investigate both structure and function of the brain and body. The major drawback of MR is its long acquisition times, which can easily exceed 30 minutes per subject scanned (Zbontar et al., 2018). These long acquisition times make MR susceptible to motion artifacts and difficult or impossible to image dynamic physiology.

MR data is collected in Fourier domain, known as k-space, and acquisition times are proportional to k-space sampling rates. Compressed sensing (CS) MR reconstruction is a technique that reconstructs high quality images from MR data incoherently sampled at rates inferior to the Nyquist-Shannon sampling theorem (Lustig et al., 2008).

In recent years, several deep-learning-based MR compressed sensing reconstruction techniques have been proposed (Zhang et al., 2018; Seitzer et al., 2018; Jin et al., 2017; Lee et al., 2017; Quan et al., 2018; Schlemper et al., 2018a; Yang et al., 2018; Zhu et al., 2018; Souza and Frayne,

2018; Eo et al., 2018b,a; Schlemper et al., 2018b; Yu et al., 2017). This rapid growth in number of publications can be explained by the success of deep learning in many medical imaging problems (Litjens et al., 2017) and its potential to reconstruct images in real-time. Traditional CS methods are iterative and usually are not suitable for fast reconstruction.

In this work, a hybrid frequency-domain/image-domain cascade of convolutional neural networks (CNNs) trained end-to-end for MR CS reconstruction is proposed. We analyze it on a single-coil acquisition setting, since many older scanners still use it (Zbontar et al., 2018), and it also works as a proof of concept that can potentially be generalized to more complex scenarios, such as parallel imaging (Deshmane et al., 2012). We compare our method with five recently published deep-learning-based models using MR raw data. We tested our model with acceleration factors of  $4\times$  and  $5\times$  (corresponding to reductions in data acquisition of 75% and 80%, respectively). Our results indicate that the improvements in our hybrid cascade are statistically significant compared to five other approaches.(Yang et al., 2018; Quan et al., 2018; Souza and Frayne, 2018; Schlemper et al., 2018a; Eo et al., 2018a)

## 2. Brief Literature Review

In the past couple years, many deep-learning models were proposed for MR CS reconstruction. Most of them were validated using private datasets and a single-coil acquisition setting. Initial works on MR CS reconstruction proposed to use U-net (Ronneberger et al., 2015) architectures with residual connections (Jin et al., 2017; Lee et al., 2017) to map from zero-filled k-space aliased reconstructions to unaliased reconstructions. Yu et al. (2017) proposed a deep de-aliasing network that incorporated a perceptual and an adversarial component. Their work was further enhanced by Yang et al. (2018). They proposed a deep de-aliasing generative adversarial network (DAGAN) that uses a residual U-net as its generator combined with a loss function that incorporates image domain, frequency domain, perceptual and adversarial information. Quan et al. (2018) proposed a generative adversarial network with a cyclic loss (Zhu et al., 2017). The cyclic loss tries to enforce that the mapping between input and output is a bijection, *i.e.* invertible.

The work of Schlemper et al. (2018a) moved away from U-nets. They proposed and implemented a model that consists of a deep cascade (Deep-Cascade) of CNNs intercalated with data consistency (DC) blocks that replace the network estimated k-space frequencies by frequencies obtained in the sampling process. Seitzer et al. (2018) built upon Deep-Cascade by adding a visual refinement network that is trained independently using the result of Deep-Cascade as its input. In their experiments, their results improved in terms of semantic interpretability and mean opinion scores, but Deep-Cascade was still better in terms of peak signal to noise ratio (PSNR). Schlemper et al. (2018b) incorporated dilated convolutions and a stochastic component on the Deep-Cascade model. All techniques discussed so far use the aliased zero-filled reconstruction as a starting point. Frequency domain information is only used either in the network loss function computation (*e.g.*, DAGAN) or in the DC blocks to recover the sampled frequencies.

Zhu et al. (2017) proposed a unified framework for reconstruction called automated transform by manifold approximation (AUTOMAP). It tries to learn the transform from undersampled k-space to image domain through fully connected layers followed by convolutional layers in image domain. The major drawback of their proposal is that their parameter complexity grows quadratically with the number of image pixels (voxels). For  $256 \times 256$  images, AUTOMAP model has  $> 10^{10}$  trainable parameters. Eo et al. (2018a) proposed a dual domain architecture that cascades k-space domain

networks with image domain networks interleaved by data consistency layers and the appropriate domain transform. The major advantage of KIKI-net is that it does not try to learn the domain transform. Therefore, it reduces the number of trainable parameters compared to AUTOMAP by a factor of 10,000, while still leveraging information from both k-space and image domains. In their model, each of the four networks that compose KIKI-net is trained independently. KIKI-net is the deepest model proposed so far for MR CS, it has one hundred convolutional layers and reconstruction time of a single  $256 \times 256$  slice is of 14 seconds on a NVIDIA GeForce GTX TITAN graphics processing unit (GPU), which is prohibitive for real time reconstruction.

[Souza and Frayne \(2018\)](#) proposed the W-net model, which consists of a k-space U-net connected to an image domain U-net through the inverse Fourier Transform (FT). The W-net model is trained end-to-end as opposed to KIKI-net and it also does not try to learn the domain transform. W-net reconstructions were shown to arguably work better (*i.e.* less process failures) with FreeSurfer ([Fischl, 2012](#)) post-processing tool.

The work of [Eo et al. \(2018b\)](#) proposes a multi-layer perceptron that estimates a target image from a one-dimensional inverse FT of k-space followed by a CNN. Their method parameter complexity grows linearly with the number of image pixels (voxels) as opposed to AUTOMAP's quadratic complexity.

Recently, the fastMRI initiative ([Zbontar et al., 2018](#)) made single-coil and multi-coil knee raw MR data publicly available for benchmarking purposes. The *Calgary-Campinas* initiative ([Souza et al., 2017](#)) has also added brain MR raw data to their dataset. Both initiatives aim to provide a standardized comparison method that will help researchers to more easily compare and assess potential improvements of new models.

### 3. Hybrid Cascade Model

Based on recent trends in the field of deep-learning-based MR reconstruction, we developed a model that incorporates elements that have improved MR reconstruction. Our proposal is a hybrid unrolled cascade structure with DC layers in between consecutive CNN blocks that is fully trained end-to-end (Figure 1). We opted not to use an adversarial component in our model for two main reasons: 1) The model already outputs realistic images that are hard for an human expert to tell apart from a fully sampled reconstruction (see results); 2) The discriminator block can always be incorporated subsequently to the reconstruction (*i.e.*, generator) network training.

Our hybrid cascade model receives as input the zero-filled reconstruction from undersampled k-space, which is represented as a two channel image. One channel stores the real part and the other stores the imaginary part of the complex number.

The first CNN block in the cascade, unlike KIKI-net, is an image domain CNN. The reason for this is that k-space is usually heavily undersampled at higher spatial frequencies. If the cascade started with a k-space CNN block, there would potentially be regions where the convolutional kernel would have no signal to operate upon. Thus, a deeper network having a larger receptive field would be needed, which would increase reconstruction times. By starting with an image domain CNN block and because of the global property of the FT, the output of this network has a corresponding k-space that is now complete. This allows the subsequent CNN block, which is in k-space domain, to perform better due to the absence region without signal (*i.e.*, because of zero-filling) without the necessity of making the network deeper. The last CNN block of our architecture is also in image domain. This decision was made empirically. Between the initial and final image domain CNN

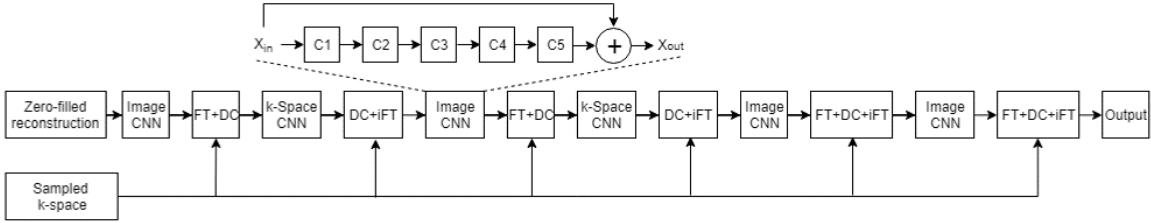


Figure 1: Architecture of the proposed Hybrid Cascade model. It has four image domain CNN blocks and two k-space CNN blocks. We start and end the network using an image domain CNN.

blocks, we alternate between k-space and image domain CNN blocks. Connecting the CNN blocks, we have the appropriate domain transform (FT or inverse FT) and the DC operator. It is important to emphasize that unlike AUTOMAP, we do not learn the FT.

Our CNN block architecture is independent of operating in the k-space or image domains. It is a residual network with five convolutional layers. The first four layers have 48 convolutional filters with  $3 \times 3$  kernels. The activations are leaky rectifier linear units with  $\alpha = 0.1$ . The final layer has two convolutional filters with a  $3 \times 3$  kernel size followed by a linear activation. All convolutional layers include bias terms. This architecture was empirically determined.

Our hybrid cascade architecture (Figure 1) has a total of four image domain and two k-space domain CNN blocks. We train it using the mean squared error cost function. It has 380,172 trainable parameters, which is relatively small compared to other deep learning architectures, such as the U-net that has  $> 20,000,000$  parameters. The number of parameters of our hybrid cascade model is in the same order of magnitude as Deep-Cascade ( $\sim 500,000$ ) and KIKI-net ( $> 3.5$  million) architectures. The main difference versus Deep-Cascade is our dual domain component. The distinction to KIKI-net is that our network is trained end-to-end and the hybrid cascade starts operating in the image domain as opposed to the k-space domain, allowing our network to have fewer layers and consequently being able to reconstruct images faster.

The depth of our cascade was experimentally set. Our source code will be made public available at <https://github.com/rmsouza01/CD-Deep-Cascade-MR-Reconstruction>. It allows you to experiment with other cascade depths, CNN depths and select the domain of each CNN block.

## 4. Experimental setup

### 4.1. Dataset

We use the *Calgary-Campinas* brain MR raw data in this work (<https://sites.google.com/view/calgary-campinas-dataset/home>). The dataset has 45 volumetric T1-weighted fully sampled k-space datasets acquired on a clinical MR scanner (Discovery MR750; General Electric (GE) Healthcare, Waukesha, WI). The data was acquired with a 12-channel imaging coil, which was combined to simulate a single-coil acquisition using vendor supplied tools (Orchestra Toolbox; GE Healthcare). The inverse FT was applied in one dimension and Gaussian 2-dimensional sampling was performed retrospectively on the other two dimensions. Our training set has 4,254 slices

coming from 25 subjects. The validation and test sets have 1,700 slices each corresponding to the remaining 20 subjects. These train, validation and test slices come from a disjoint set of subjects.

#### 4.2. Metrics and statistical analysis

The metrics used in this work were normalized mean squared error (NRMSE), PSNR and Structural Similarity (SSIM) (Wang et al., 2004). Low NRMSE and high PSNR and SSIM values represent good reconstructions. The metrics are computed against the fully sampled reconstruction. These metrics were chosen seeing that they are commonly used to assess CS MR reconstruction. We assessed statistical significance using paired Wilcoxon signed-rank test with an alpha of 0.05.

#### 4.3. Compared Methods

We compared our method, which we will refer to as Hybrid-Cascade, against four previously published deep-learning-based methods that had publicly available source code and our own implementation of KIKI-net, which we will refer as KIKI-net-like. It has 6 CNN blocks alternating between frequency-domain and image-domain CNNs interleaved by DC blocks. Our KIKI-net-like implementation has the same number of trainable parameters as Hybrid-Cascade. Our goal, when comparing to KIKI-net-like, is to gain empirical evidence that initiating the cascade on image-domain can potentially lead to better reconstructions. The compared methods with public source code were: DAGAN (Yang et al., 2018), RefineGAN (Quan et al., 2018), W-net (Souza and Frayne, 2018) and Deep-Cascade (Schlemper et al., 2018a).

All networks were re-trained from scratch for two different sampling rates: 25% and 20% corresponding to speed-ups of  $4\times$  and  $5\times$ , respectively. We used fixed Gaussian sampling patterns throughout training and testing (Figure 2).

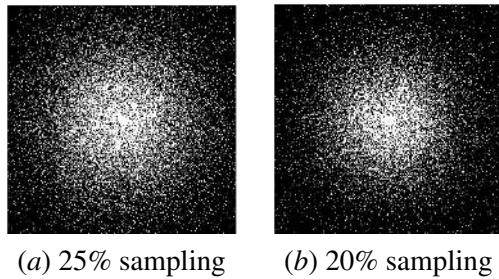


Figure 2: Gaussian sampling patterns used in the experiments.

### 5. Results and Discussion

Hybrid-Cascade was the top performing method for all metrics and acceleration factors. Although Hybrid-Cascade results were close to Deep-Cascade and KIKI-net-like, the difference was statistically significant for NRMSE and PSNR ( $p < 0.05$ ). DAGAN and RefineGAN achieved the poorer results. W-net was ranked fourth best. Quantitative results are summarized in Table 1.

DAGAN and RefineGAN lose relevant brain structural information in their reconstructions. W-net outputs visually pleasing reconstructions, but it lacks textural information which is encoded in

the high frequencies. Hybrid-Cascade, Deep-Cascade and KIKI-net-like output very similar reconstructions, but small differences can be noticed specially in the cerebellum region. Sample reconstructions for each technique are depicted in Figure 3. Starting with an image-domain CNN lead to a higher error reduction in the first block of the cascade as opposed to starting with a k-space CNN (Figure 4).

It is interesting to notice that the top tree techniques in our analysis, Hybrid-Cascade, KIKI-net-like and Deep-Cascade all use unrolled structures combined with DC. DAGAN, RefineGAN and W-net all use some variation of a U-net architecture within their models. This make us conjecture that flat unrolled CNN architectures may be better suited models for MR CS reconstruction.

Table 1: Summary of the results for the different architectures and different acceleration factors. The best results for each metric and acceleration factor are emboldened.

Acceleration factor	Model	NRMSE (%)	PSNR (dB)	SSIM
$4\times$	DAGAN	$2.925 \pm 1.474$	$31.330 \pm 3.112$	$0.84 \pm 0.11$
	RefineGAN	$1.898 \pm 1.300$	$35.436 \pm 3.705$	$0.90 \pm 0.07$
	W-net	$1.364 \pm 1.011$	$38.228 \pm 3.317$	$0.93 \pm 0.07$
	KIKI-net-like	$1.178 \pm 1.022$	$39.640 \pm 3.355$	$0.95 \pm 0.06$
	Deep-Cascade	$1.198 \pm 1.057$	$39.510 \pm 3.345$	$0.95 \pm 0.07$
	Hybrid-Cascade	$1.151 \pm 1.022$	$39.871 \pm 3.380$	$0.96 \pm 0.06$
$5\times$	DAGAN	$3.866 \pm 1.435$	$28.691 \pm 2.658$	$0.79 \pm 0.11$
	RefineGAN	$2.273 \pm 1.401$	$33.844 \pm 3.825$	$0.87 \pm 0.09$
	W-net	$1.645 \pm 1.085$	$36.501 \pm 3.226$	$0.92 \pm 0.09$
	KIKI-net-like	$1.452 \pm 1.092$	$37.669 \pm 3.224$	$0.94 \pm 0.08$
	Deep-Cascade	$1.453 \pm 1.106$	$37.668 \pm 3.202$	$0.94 \pm 0.08$
	Hybrid-Cascade	$1.423 \pm 1.099$	$37.875 \pm 3.252$	$0.94 \pm 0.08$

Intermediary outputs of Hybrid-Cascade in a sample subject are depicted in Figure 5. The input zero-filled reconstruction has a NRMSE of 14.76% and it drops to 2.41% after the first CNN block, which is the largest error drop throughout the cascade. The error keeps lowering consistently up to the fifth CNN block. Then, the error goes up, but it immediately goes back down again in the final CNN block. This finding was consistent across all test slices. Although an odd finding, it is not unexpected. Since the network was optimized to minimize the mean squared error of the final CNN block output, the error across intermediary outputs can potentially oscillate as it happened in this case.

Concerning reconstruction times, we did not perform a systematic assessment. Our Hybrid-Cascade and KIKI-net-like implementations take on average 22 milliseconds to reconstruct a  $256 \times 256$  slice on a NVIDIA GTX 1070 GPU, which is considerably faster than the 14 seconds that the original KIKI-net proposal takes to reconstruct a same size slice. W-net, DAGAN and RefineGAN also have reconstructions times in the order of a few milliseconds.

The Hybrid-Cascade model can be applied to multi-coil reconstruction by processing each coil k-space independently, and then combining the resulting images through a sum of squares algorithm. This approach would probably not be optimal, as it would disregard complementary information across the k-spaces from each coil. The extension of DC to multi-coil data is not straightforward and is still an open research question.

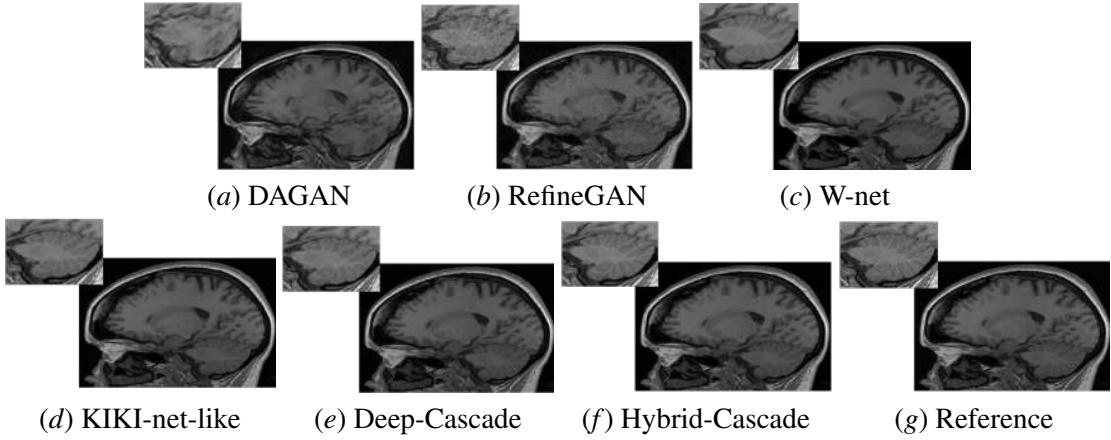


Figure 3: Sample reconstructions for all different reconstruction techniques assessed using a speed-up factor of  $5\times$ . Visually, Hybrid-Cascade, Deep-Cascade and KIKI-net-like are the most similar to the fully sampled reconstruction reference. W-net also presents a visually pleasing reconstruction, but it lacks textural information, i.e. high frequencies information is attenuated.

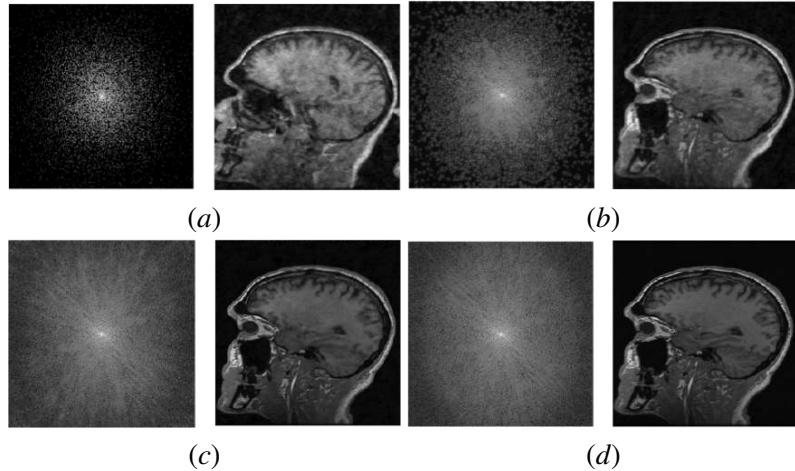


Figure 4: (a) Undersampled k-space and its corresponding zero-filled reconstruction (NRMSE=15.2%). (b) Output of first CNN block of KIKI-net-like (NRMSE=4.0%). (c) Output of first CNN block of Hybrid-Cascade (NRMSE=2.5%) and (d) reference fully sampled k-space and its image reconstruction.

## 6. Conclusions

We proposed a hybrid frequency-domain/image-domain cascade of CNNs for MR CS reconstruction. We compared it with the current state-of-the-art of deep-learning-based reconstructions using a public dataset. The differences between our model and the compared ones were statistically significant ( $p < 0.05$ ).

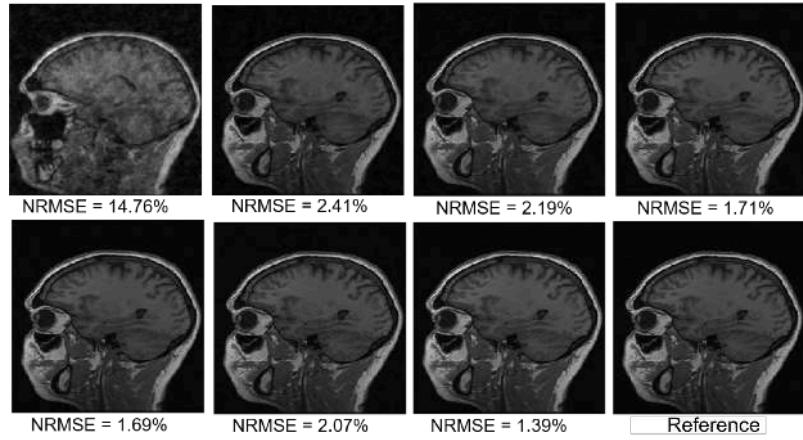


Figure 5: From the top left to the bottom right: input zero-filled reconstruction for a speed-up factor of  $5\times$ , output of each CNN block in the Hybrid-Cascade, and reference fully sampled reconstruction. An interesting finding is that the NRMSE increases at the output of the fifth CNN block in the cascade and than it decreases again after the sixth block. These finding was consistent across all slices in the test set.

As future work, we intend to investigate how to adapt our model to parallel imaging combined with CS using the full spectrum of information across the coils. We also would like to explore how our dual domain model potentially relates to commonly used parallel imaging methods, such as Generalized Autocalibrating Partially Parallel Acquisitions (GRAPPA) (Griswold et al., 2002), which works on k-space domain, and Sensitivity Encoding for fast MR imaging (SENSE) (Pruessmann et al., 1999), which works on image domain.

## Acknowledgments

The authors would like to thank the Natural Science and Engineering Council of Canada (NSERC) for operating support, NVidia for providing a Titan V GPU, and Amazon Web Services for access to cloud-based GPU services. We would also like to thank Dr. Louis Lauzon for setting up the script to save the raw MR data at the Seaman Family MR Centre. R.S. was supported by an NSERC CREATE I3T Award and currently holds the T. Chen Fong Fellowship in Medical Imaging from the University of Calgary. R.F. holds the Hopewell Professorship of Brain Imaging at the University of Calgary.

## References

- Anagha Deshmane, Vikas Gulani, Mark A Griswold, and Nicole Seiberlich. Parallel mr imaging. *Journal of Magnetic Resonance Imaging*, 36(1):55–72, 2012. ISSN 1522-2586.
- Taejoon Eo, Yohan Jun, Taeseong Kim, Jinseong Jang, Ho-Joon Lee, and Dosik Hwang. Kikinet: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. *Magnetic resonance in medicine*, 2018a. ISSN 0740-3194.

Taejoon Eo, Hyungseob Shin, Taeseong Kim, Yohan Jun, and Dosik Hwang. Translation of 1d inverse fourier transform of k-space to an image based on deep learning for accelerating magnetic resonance imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 241–249. Springer, 2018b.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase. Generalized autocalibrating partially parallel acquisitions (grappa). *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002. ISSN 0740-3194.

Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017. ISSN 1057-7149.

Dongwook Lee, Jaejun Yoo, and Jong Chul Ye. Deep residual learning for compressed sensing MRI. In *IEEE 14th International Symposium on Biomedical Imaging*, pages 15–18. IEEE, 2017. ISBN 1509011722.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing MRI. *IEEE signal processing magazine*, 25(2):72–82, 2008. ISSN 1053-5888.

Dwight G Nishimura. *Principles of magnetic resonance imaging*. Stanford Univ., 1996.

Klaas P Pruessmann, Markus Weiger, Markus B Scheidegger, and Peter Boesiger. Sense: sensitivity encoding for fast mri. *Magnetic resonance in medicine*, 42(5):952–962, 1999. ISSN 1522-2594.

Tran Minh Quan, Thanh Nguyen-Duc, and Won-Ki Jeong. Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss. *IEEE transactions on medical imaging*, 37(6):1488–1497, 2018. ISSN 0278-0062.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503, 2018a. ISSN 0278-0062.

Jo Schlemper, Guang Yang, Pedro Ferreira, Andrew C. Scott, Laura-Ann McGill, Zohya Khalique, Margarita Gorodezky, Malte Roehl, Jennifer Keegan, Dudley Pennell, David N. Firmin, and Daniel Rueckert. Stochastic deep compressive sensing for the reconstruction of diffusion tensor cardiac mri. In *MICCAI*, 2018b.

Maximilian Seitzer, Guang Yang, Jo Schlemper, Ozan Oktay, Tobias Würfl, Vincent Christlein, Tom Wong, Raad Mohiaddin, David Firmin, and Jennifer Keegan. Adversarial and perceptual refinement for compressed sensing mri reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–240. Springer, 2018.

Roberto Souza and Richard Frayne. A hybrid frequency-domain/image-domain deep network for magnetic resonance image reconstruction. *arXiv preprint arXiv:1810.12473*, 2018.

Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Letícia Rittner, Richard Frayne, and Roberto Lotufo. An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage*, 2017. ISSN 1053-8119.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. ISSN 1057-7149.

Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, and Yike Guo. DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE transactions on medical imaging*, 37(6):1310–1321, 2018. ISSN 0278-0062.

Simiao Yu, Hao Dong, Guang Yang, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, and David Firmin. Deep de-aliasing for fast compressive sensing mri. *arXiv preprint arXiv:1705.07137*, 2017.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

Pengyue Zhang, Fusheng Wang, Wei Xu, and Yu Li. Multi-channel generative adversarial network for parallel magnetic resonance image reconstruction in k-space. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 180–188. Springer, 2018.

Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487, 2018. ISSN 1476-4687.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.

# 3D multirater RCNN for multimodal multiclass detection and characterisation of extremely small objects

**Carole H. Sudre<sup>1,2,3</sup>**

CAROLE.SUDRE@KCL.AC.UK

**Beatrix Gomez Anson<sup>4</sup>**

BGOMEZA@SANTPAU.CAT

**Silvia Ingala<sup>5</sup>**

S.INGALA@VUMC.NL

**Chris D. Lane<sup>2</sup>**

C.LANE@UCL.AC.UK

**Daniel Jimenez<sup>2</sup>**

D.JIMENEZ@UCL.AC.UK

**Lukas Haider<sup>6</sup>**

L.HAIDER@UCL.AC.UK

**Thomas Varsavsky<sup>1,3</sup>**

THOMAS.VARSAVSKY@KCL.AC.UK

**Lorna Smith<sup>7</sup>**

LORNA.SMITH@UCL.AC.UK

**Sébastien Ourselin<sup>1</sup>**

SEBASTIEN.OURSELIN@KCL.AC.UK

**Rolf H Jäger<sup>8</sup>**

R.JAGER@UCL.AC.UK

**M. Jorge Cardoso<sup>1,2,3</sup>**

M.JORGE.CARDOSO@KCL.AC.UK

<sup>1</sup> School of Biomedical Engineering and Imaging Sciences, King’s College London, UK

<sup>2</sup> Dementia Research Centre, UCL Institute of Neurology, UK

<sup>3</sup> Department of Medical Physics and Biomedical Engineering, University College London, UK

<sup>4</sup> Santa Creu i Sant Pau Hospital, Universitat Autònoma Barcelona, Barcelona, Spain

<sup>5</sup> Vrije University Medical Centre Amsterdam, The Netherlands

<sup>6</sup> Queen Square Multiple Sclerosis Centre, UCL Institute of Neurology, London, UK

<sup>7</sup> Cardiometabolic Phenotyping Group, Institute of Cardiovascular Science, UCL, London, UK

<sup>8</sup> Brain Repair and Rehabilitation Group, Institute of Neurology, UCL, London, UK

## Abstract

Extremely small objects (ESO) have become observable on clinical routine magnetic resonance imaging acquisitions, thanks to a reduction in acquisition time at higher resolution. Despite their small size (usually <10 voxels per object for an image of more than  $10^6$  voxels), these markers reflect tissue damage and need to be accounted for to investigate the complete phenotype of complex pathological pathways. In addition to their very small size, variability in shape and appearance leads to high labelling variability across human raters, resulting in a very noisy gold standard. Such objects are notably present in the context of cerebral small vessel disease where enlarged perivascular spaces and lacunes, commonly observed in the ageing population, are thought to be associated with acceleration of cognitive decline and risk of dementia onset. In this work, we redesign the RCNN model to scale to 3D data, and to jointly detect and characterise these important markers of age-related neurovascular changes. We also propose training strategies enforcing the detection of extremely small objects, ensuring a tractable and stable training process.

## 1. Introduction

The vascular network that supplies the brain changes with age, inducing alterations to surrounding tissue. Macroscopic changes, hallmark of cerebral small vessel disease, can be observed on structural MR images and include white matter hyperintensities, lacunar infarcts, cerebral micro-

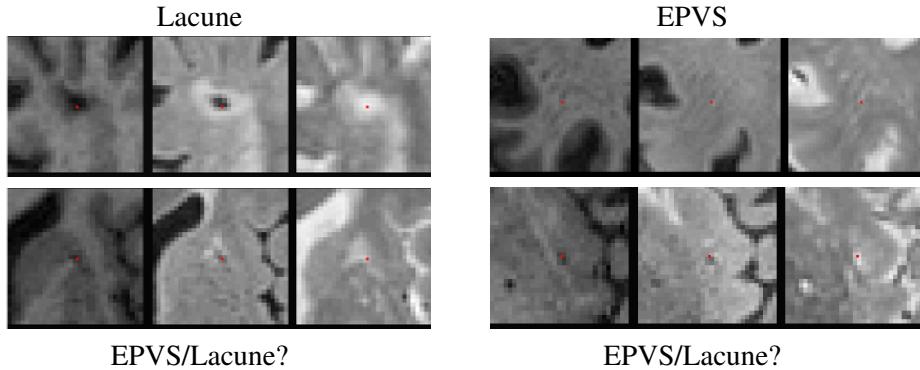


Figure 1: Examples of EPVS and lacunes on which agreement was high (top row) or low (bottom row) on the three structural modalities of interest (T1, FLAIR, T2). The red dot indicates the centre of mass of the object of interest.

haemorrhages and enlarged perivascular spaces (EPVS), among others (Wardlaw et al., 2013). More specifically, perivascular spaces are thought to be used as a lymphatic pathway in a drainage mechanism, where entrapped fluid can extend this space, making it visible in MR images, often as linearly-shaped fluid-like structure (Figure 1 top right). In clinical practice, their presence is classically assessed using visual scales on T2 MR images, described as elongated bright ellipsoids (Potter et al., 2015). The use of such visual scales requires extensive training and expertise, is prone to inter/intra rater variability, suffers from flooring/ceiling effects and is time-consuming for the operator. Some works have recently been proposed to automatically assess the EPVS burden (Boespflug et al., 2018)(Dubost et al., 2019) in clinical grade MR data, while others propose to segment EPVS at higher field (7T) (Zhang et al., 2016). In contrast, lacunar infarcts, observed with a much lower frequency, are areas of dead tissue due to complete ischemia. Their shape signature is an ovoid object of 3 – 15mm of diameter, with a cerebrospinal fluid (CSF) -like intensity in the centre. Often, on T2 weighted Fluid attenuated inversion recovery (FLAIR) images, they are surrounded by a rim of hyperintensity (see Figure 1 top left). In practice, even for trained radiologists, distinguishing between EPVS and lacunes can be very challenging (see Figure 1 bottom). This results in double counting of uncertain objects (del C. Valdés Hernández et al., 2013), and under-counting when objects branch from the same point. This task is however of clinical importance as these markers reflect tissue damage and are key to understand complex pathological pathways of age-related vascular changes (Ramirez et al., 2016).

To account for the above-mentioned challenges, we propose to adapt the 2D RCNN model presented by He et al (He et al., 2017) that allows for multiclass multi-instances simultaneous detection and segmentation to multirater 3D data, in the context of EPVS and lacune detection and size characterisation, with the perspective of a future expansion to more object classes (e.g. white matter hyperintensities) and their semantic segmentation. After a brief description of the 2D RCNN framework, we detail the challenges inherent to 3D data of such a framework in the capture of extremely small objects, and describe the introduction of multirater predictions.

## 2. Methods

### 2.1. Two dimensional RCNN

In the original RCNN framework, a backbone network is trained to extract generic features. This initial training is then complemented by two stages: a region proposal network and a final classification network applied to selected boxes whose shapes have been modified to fit a specified mask. In the 2D setting, the region proposal network is based on the classification as positive or negative of a series of predefined boxes created based on anchors, regularly spaced on the 2D grid with different ratios of height and width. All selected grid are then resampled (pooled) to a user-specified shape and fed to the final segmentation classification branch of the framework.

### 2.2. Challenges and strategies for a multirater 3D extension

The main challenges related to the extension of the successful RCNN framework to 3D data lay in the memory and data requirements, as well as an extreme class imbalance. In terms of memory, the generation of grid anchors become notably prohibitive in 3D. Additionally, when dealing with ESOs, any interpolation induced by the region pooling may obscure relevant features and render the segmentation meaningless. In order to account for these challenges, the following strategies were adopted at the different stages of the framework:

**Backbone network** The 3D HighResNet proposed by Li et al. ([Li et al., 2017](#)) was used as backbone network to extract features. This architecture has a large contextual field of view at reduced parameter cost. This network uses three levels of residual convolutional networks with dilated convolutions with increasing dilation factor, each level consisting of three dilated convolutions with fixed dilation factor alternating with batch normalisation and ReLu activation. In the presented setting, the network was applied to regress a distance map with a root mean square error loss. The distance map is calculated from each given element's segmentation.

**Region Proposal Network (RPN)** In order to alleviate the memory burden of having to explicitly describe anchors and associated boxes, the RPN, consisting of one classification and one regression branch, was applied in a convolutional fashion to every voxel. The features extracted at the backbone level were fed into a small convolutional network with a single common  $3^3$  kernel, followed by either a classification layer or a regression layer. The classification layer establishes if the centre of the patch is likely to be the centre of mass of the target object, while the regression part outputs four values: three values representing the distance to the closest object centre of mass, and the fourth representing the scale of the targeted object. Classification and regression were learnt from 300 samples on the patch, with a 50/50 balance between positive and negative samples. To avoid any impact on the regression branch, negative samples did not bear any weight on the regression loss. A cross-entropy loss was used for the classification branch while a smooth distance loss was applied on the regression branch for the estimation of the distance to the closest element centre of mass. Denoting  $r_n$  the absolute error between predicted value and ground truth for a given sample  $n$ , the smooth distance loss  $DL$  is expressed as:

$$DL = \frac{1}{N} \sum_{n=1}^N f(r_n) \text{ where } f(r_n) = \begin{cases} 0.5r_n^2 & \text{if } r_n < 0.5 \\ (r_n - 0.125)^2 - 2 & \text{if } r_n > 2.125 \\ r_n - 0.125 & \text{otherwise} \end{cases}$$

**Refinement/Classification Network (RCN)** From the location of proposed ESO centres-of-mass, boxes were associated with ground-truth objects, and extracted masks are directly fed so as to classify the boxes and adjust the regression of the centre of mass.

The branch jointly classifying the element and regressing centre of mass and object scale consisted of a convolutional layer of kernel size 7, followed by a fully connected layer. After average pooling, classification and bounding box regression were established. For the regression branch, the target prediction was the residual between the RPN prediction and the ground truth for the three location elements, and a scale correction factor for the size. A similar smooth distance loss was applied as a cost function. In contrast to the original RCNN framework, selected boxes were neither resized nor pooled to a predefined shape. This is in order to avoid interpolation that would be detrimental, given that many of the targeted elements are one voxel wide.

**Multirater encoding** For each of the manually-segmented elements, the raters were asked to attribute one of the following class: 1)Nothing; 2) Lacune; 3) EPVS; 4) Undecided between lacune and EPVS; Instead of a crisp classification, a soft probability label was obtained as the average of the multiple raters involved in the classification and used as target. For each rater, a fully connected layer was added in order to directly infer the classification of each individual. The architecture framework is displayed in Figure 2.

### 2.3. Implementation

**Sampling and data normalisation** The existence of two types of imbalance (foreground vs background, and between EPVS vs lacunes) required a purpose-specific sampling scheme. A probabilistic weight sampling was adopted as suggested by Ronneberger et al (Ronneberger et al., 2015) to extract patches of size  $64^3$  over the images. For this purpose, the inverse of the distance maps from segmented EPVS and lacunes were smoothed and linearly combined using a ratio of 1/100 reflecting the relative frequency of occurrence of these two classes. These maps were clipped to

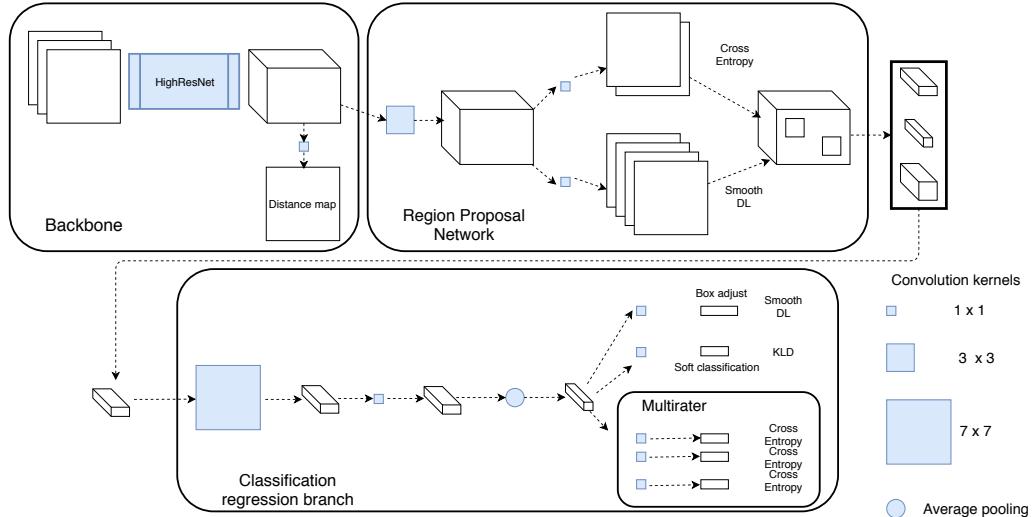


Figure 2: Architecture of the 3D multirater RCNN for extremely small objects.

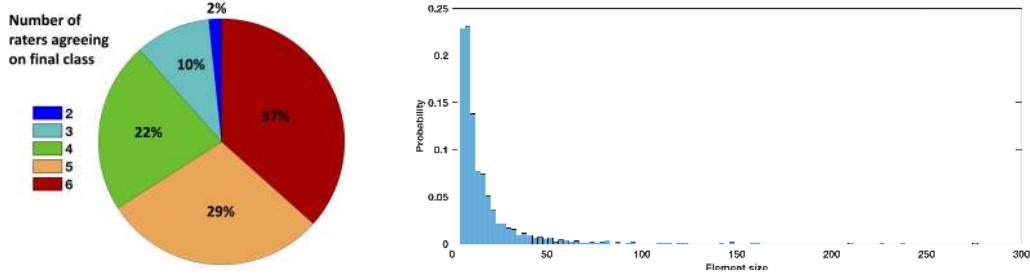


Figure 3: Repartition of agreement between raters responsible for the final crisp classification (left) and distribution of the size of the targeted elements (right).

a minimum of  $10^{-5}$  to reflect the overall background/foreground ratio. All input data (T1, T2 and FLAIR images) was bias field corrected, skull stripped, and then z-scored to the white matter region statistics, segmented independently ([Sudre et al., 2015](#)).

**Training scheduling and loss functions** The framework was implemented within NiftyNet ([Gibson et al., 2018](#)) ([niftynet.io](#)) and will be merged into the main codebase at the time of publication. The network was trained progressively per stage to mitigate training stability issues. Sections where classification and regression were combined (RPN and RCN) were trained in two steps: the first one consisted of the classification training with a sigmoid applied to the regression loss, and the second step was the sum of the two losses. Each of the steps was trained for 1000 iterations with learning rate of 0.0001. In order to account for scale differences observed across combined loss functions, notably between classification and positioning regression losses, empirical weights were chosen and progressively modified throughout the training of the network in order to always ensure a balance between classification accuracy and box positioning.

**Inference** At inference, a similar patch size was used as for the training step in order to expect a similar number of proposals (limited to 300). In order to prune the potential positions of centre of mass, the information from the score map and the distance map were combined. The score map was thresholded at 0.25 and the morphological skeleton of the underlying distance map were extracted. The corresponding distance score maxima were then taken as potential proposed centres of mass. Centres of mass closer than 2mm were pruned as a form of non-maximum suppression.

### 3. Data and experiments

#### 3.1. Data

16 subjects were selected out of a longitudinal tri-ethnic cohort of elderly subjects aiming at investigating the relationship between cardiovascular risk factors and brain health ([Tillin et al., 2012](#)). At the third wave of investigation, subjects of this cohort underwent an MR session including the acquisition of 3D 1mm<sup>3</sup> isotropic T1 weighted, T2 weighted and T2-weighted FLAIR images ([Sudre et al., 2018](#)). The 16 subjects were chosen for their elevated vascular burden visually assessed by a trained radiologist. EPVS and lacunes were manually segmented on the three available structural MR sequences using ITKSnap ([Yushkevich et al., 2006](#)). Performed by a rater accustomed to the use of the segmentation software, the delineation of EPVS and lacunes for a single subject required

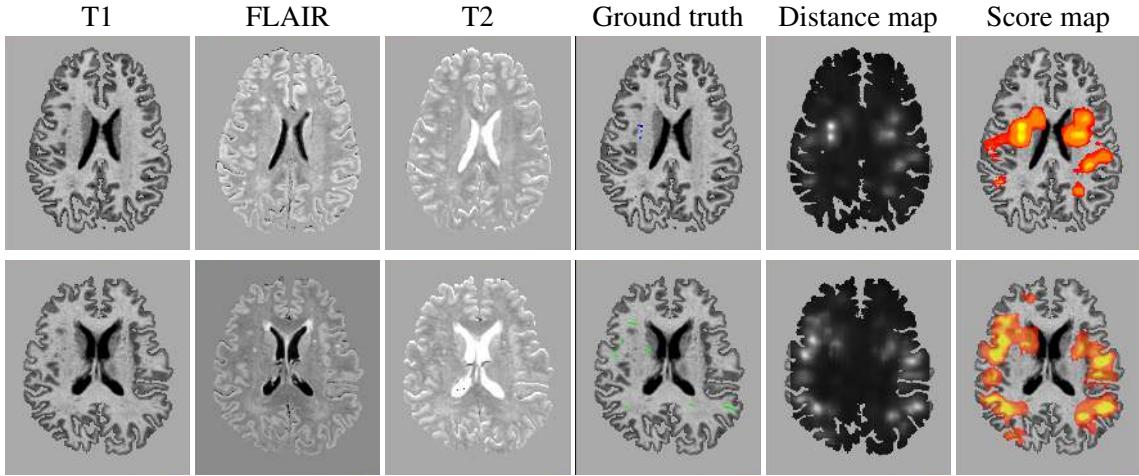


Figure 4: Two holdout cases with the three input channels (T1, FLAIR, T2), gold standard segmentation, inferred distance maps and score map.

an average of 20h. Segmentations were done in a multi-view manner to ensure geometrical consistency, with all images aligned to the T1 sequence as a geometrical reference. Segmentation masks were then automatically corrected and voxels with inappropriate signal identity signature were removed. Individual EPVS and lacunes were further classified at the level of connected components by six operators with a varied range of expertise using an in house dedicated viewer. Only elements with a volume of more than 5 voxels were considered in this work, resulting in a database of 2442 elements. The volumes of segmented elements ranged thus from 5 to 350, with 48.8% with a size below 10 voxels. Perfect agreement among raters was reached only in 36.6% of the cases, and only 2.8% of the elements were ultimately classified as lacunes. Figure 3 presents an histogram of element size, and a pie chart representing the proportion or rater agreement. The poor inter-rater classification agreement hints at the complexity of the task. Uncertainty over the segmentation would have to be evaluated over multiple raters before envisioning moving the proposed object RCNN detection model to a full Mask-RCNN, also performing segmentation. Due to the lack of more training data, 14 of the subjects were used for training and 2 were hold-out for testing.

### 3.2. Experiments

In order to compare the performance of a standard segmentation approach to the proposed multi-class detection framework, we trained semantic segmentation models with multiple combinations of architectures, loss functions (e.g Generalised Dice Loss), learning rates (from  $10^{-6}$  to  $10^{-3}$ ) and regularisation. Parameter choice was similar to the one used for the backbone network, with ranges that have been shown to perform well on unbalanced data. Unfortunately, no network was able to segment any foreground class.

We present hereafter the results obtained at the different stages of the model in terms of distance regression, score map, RPN and multirater classification.

## 4. Results

Each step of the framework was assessed on the two held out test subjects using the same metrics as the loss functions. Figure 4 presents the input data for the three modalities along with the ground truth segmentation, the regressed distance map and the inferred score map.

Interestingly, some elements not present in the gold standard segmentation but detected as per the score map were a posteriori considered as valid enlarged perivascular spaces as can be seen on Figure 5.

Given the limitations of the available gold standard in terms of inter-rater element classification, and potential missing objects, the validation focused on the sensitivity of the trained model and the relationship of the results with the multi-rater uncertainty. A sensitivity of 72.7% was observed across the two test subjects with a significant difference in element size between false negatives and true positives (Wilcoxon ranksum test  $p < 0.00001$ ). Investigating the relationship between the ratio of overlap between best matching detected box and ground truth proposal, a significant association between agreement of raters and overlap was observed ( $p=0.002$ ) with a median overlap of 59% when all raters agreed and an overlap of 30% for the more uncertain cases (at least one rater considering the element not to be relevant). Note that overlap is measured on the predicted box, which can vary widely in its size. Figure 6 presents boxplots of relationship between ESO scale and detection (left), and overlap ratio with rater uncertainty (right).

Figure 7 presents the ground truth and matching predicted boxes where the color reflects the probability of belonging to each of the classes (nothing - lacune - EPVS - undecided).

## 5. Discussion and conclusion

In this work we proposed a 3D deep learning model for the detection and characterisation of extremely small objects incorporating multi-rater labels and agreement. In this context, two types of extreme class imbalance were found, with a very low ratio of foreground to background, as well as a strong imbalance between the estimated classes where the prevalence of enlarged perivascular spaces being much higher than the number of lacunes.

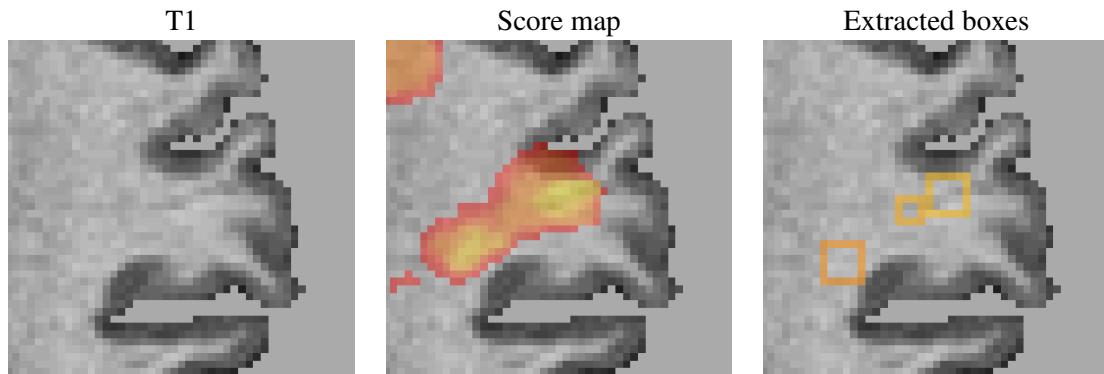


Figure 5: ESOs rightly detected by the network but missed during manual labelling. From left to right, T1, predicted score map and predicted boxes.

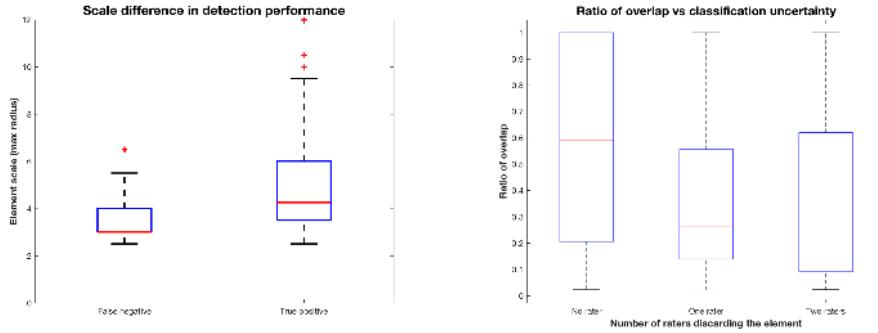


Figure 6: Left: Gold standard scale of ESO versus detection (False negative/True positive). Right: Relationship between multi-rater disagreement and box overlap performance. Note that overlap ratio is higher for more certain objects.

The different steps of the framework were evaluated, showing a good sensitivity of the region proposal network. Specificity was not ideal, probably limited by the missing annotation of individual branching elements (currently considered as a single ESO). Future work will use the multi-rater gold standard to better guide network updates by penalising classification errors made on definite classifications more strongly. Additionally, the segmentation, currently only used to obtain the original distance map, could enrich the model by defining a soft labelling at the edges and/or obtaining additional manual ratings. Furthermore, it must be noted that the training accuracy heavily depends on the quality of the initial co-registration of the different modalities, as one voxel of shift may lead to an aberrant intensity signature. At this stage, proposal boxes are cuboid, since a single scale factor is regressed at training. Future work will also involve transforming the scale regression of the

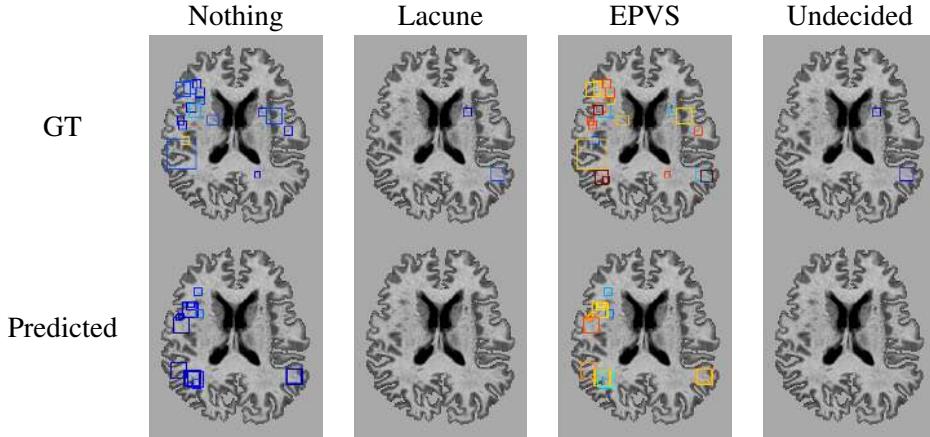


Figure 7: Probabilistic ground truth (GT) and predicted boxes for the different classes. All blue boxes correspond to low classification probabilities ( $p < 0.5$ ), and illustrate rating variability. Yellow to red boxes represent probabilities ranging from 0.5 to 1, and represent confident ESOs classifications.

RCNN into a multi direction scale factor transformation thus providing further information on the shape of the enclosed object.

## Acknowledgments

We are extremely grateful to all the participants of the SABRE study, and past and present members of the SABRE team. This work was supported by an Alzheimer's Society Junior Fellowship (AS-JF-17-011), the Wellcome/EPSRC Centre for Medical Engineering [WT 203148/Z/16/Z], IMI2 grant AMYPAD [115952], the MSCA-ITN-Demo [721820], and the Wellcome Flagship Programme in High-Dimensional Neurology. The SABRE study was funded at baseline by the Medical Research Council, Diabetes UK, and the British Heart Foundation. At follow-up, the study was funded by the Wellcome Trust (067100, 37055891 and 086676/7/08/Z), the British Heart Foundation (PG/06/145, PG/08/103/ 26133, PG/12/29/29497 and CS/13/1/30327) and Diabetes UK (13/0004774). We gratefully acknowledge NVIDIA corporation for the donation of a GPU Tesla K40 that was used in the preparation of this work.

## References

- Erin L. Boespflug, Daniel L. Schwartz, David Lahna, Jeffrey Pollock, Jeffrey J. Iliff, Jeffrey A. Kaye, William Rooney, and Lisa C. Silbert. MR Imaging-based Multimodal Autoidentification of Perivascular Spaces (mMAPS): Automated Morphologic Segmentation of Enlarged Perivascular Spaces at Clinical Field Strength. *Radiology*, 286(2):632–642, feb 2018.
- Maria del C. Valdés Hernández, Rory J. Piper, Xin Wang, Ian J. Deary, and Joanna M. Wardlaw. Towards the automatic computational assessment of enlarged perivascular spaces on brain magnetic resonance images: A systematic review. *Journal of Magnetic Resonance Imaging*, 38(4):774–785, oct 2013.
- Florian Dubost, Hieab Adams, Gerda Bortsova, M Arfan Ikram, Wiro Niessen, Meike Vernooij, and Marleen de Bruijne. 3D regression neural network for the quantification of enlarged perivascular spaces in brain MRI. *Medical image analysis*, 51:89–100, jan 2019.
- Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, et al. Niftynet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, 158:113–122, 2018.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.
- Wenqi Li, Guotai Wang, Lucas Fidon, Sébastien Ourselin, M Jorge Cardoso, and Tom Vercauteren. On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task. In *International Conference on Information Processing in Medical Imaging (IPMI)*, 2017.
- Gillian M Potter, Francesca M Chappell, Zoe Morris, and Joanna M Wardlaw. Cerebral perivascular spaces visible on magnetic resonance imaging: development of a qualitative rating scale and its observer reliability. *Cerebrovascular Diseases*, 39(3-4):224–231, jan 2015.

Joel Ramirez, Courtney Berezuk, Alicia A McNeely, Fuqiang Gao, JoAnne McLaurin, and Sandra E Black. Imaging the Perivascular Space as a Potential Biomarker of Neurovascular and Neurodegenerative Diseases. *Cellular and molecular neurobiology*, mar 2016. ISSN 1573-6830. doi: 10.1007/s10571-016-0343-6.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Carole Sudre, M Jorge Cardoso, Willem Bouvy, Geert Biessels, Josephine Barnes, and Sébastien Ourselin. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. *IEEE Transactions on Medical Imaging*, 34(10):2079–2102, apr 2015. ISSN 1558-254X. doi: 10.1109/TMI.2015.2419072.

Carole H. Sudre, Lorna Smith, David Atkinson, Nish Chaturvedi, Sébastien Ourselin, Frederik Barkhof, Alun D. Hughes, H. Rolf Jäger, and M. Jorge Cardoso. Cardiovascular Risk Factors and White Matter Hyperintensities: Difference in Susceptibility in South Asians Compared With Europeans. *Journal of the American Heart Association*, 7(21), nov 2018.

Therese Tillin, Nita G Forouhi, Paul M McKeigue, Nish for the SABRE group Chatuverdi, and Nish Chaturvedi. Southall And Brent REvisited: cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins. *International Journal of Epidemiology*, 41(1):33–42, feb 2012.

Joanna M Wardlaw, Eric E Smith, G J Biessels, Charlotte Cordonnier, Franz Fazekas, Richard Frayne, Richard I Lindley, John T O'Brien, Frederik Barkhof, Oscar R Benavente, Sandra E Black, Carol Brayne, Monique M B Breteler, Hugues Chabriat, Charles DeCarli, Frank-Erik de Leeuw, Fergus Doubal, Marco Duering, Nick C Fox, Steven Greenberg, Vladimir Hachinski, Ingo Kilimann, Vincent Mok, Robert van Oostenbrugge, Leonardo Pantoni, Oliver Speck, Blossom C M Stephan, Stefan Teipel, Viswanathan Anand, David Werring, Christopher Chen, Colin Smith, Mark A van Buchem, Bo Norrvig, Philip B Gorelick, and Martin Dichgans. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurology*, 12:822–838, 2013.

Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.

Jun Zhang, Yaozong Gao, Sang Hyun Park, Xiaopeng Zong, Weili Lin, and Dinggang Shen. Segmentation of Perivascular Spaces Using Vascular Features and Structured Random Forest from 7T MR Image. *Machine learning in medical imaging. MLMI (Workshop)*, 10019:61–68, oct 2016.

# XLSor: A Robust and Accurate Lung Segmentor on Chest X-Rays Using Criss-Cross Attention and Customized Radiorealistic Abnormalities Generation

You-Bao Tang<sup>\*1</sup>

YOUNBAO.TANG@NIH.GOV

Yu-Xing Tang<sup>\*1</sup>

YUXING.TANG@NIH.GOV

Jing Xiao<sup>2</sup>

XIAOJING661@PINGAN.COM.CN

Ronald M. Summers<sup>1</sup>

RMS@NIH.GOV

<sup>1</sup> Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892-1182, USA

<sup>2</sup> Ping An Insurance Company of China, Shenzhen, 510852, China

## Abstract

This paper proposes a novel framework for lung segmentation in chest X-rays. It consists of two key contributions, a criss-cross attention based segmentation network and radiorealistic chest X-ray image synthesis (*i.e.* a synthesized radiograph that appears anatomically realistic) for data augmentation. The criss-cross attention modules capture rich global contextual information in both horizontal and vertical directions for all the pixels thus facilitating accurate lung segmentation. To reduce the manual annotation burden and to train a robust lung segmentor that can be adapted to pathological lungs with hazy lung boundaries, an image-to-image translation module is employed to synthesize radiorealistic abnormal CXRs from the source of normal ones for data augmentation. The lung masks of synthetic abnormal CXRs are propagated from the segmentation results of their normal counterparts, and then serve as pseudo masks for robust segmentor training. In addition, we annotate 100 CXRs with lung masks on a more challenging NIH Chest X-ray dataset containing both posterioranterior and anteroposterior views for evaluation. Extensive experiments validate the robustness and effectiveness of the proposed framework. The code and data can be found from [https://github.com/rsummers11/CADLab/tree/master/Lung\\_Segmentation\\_XLSor](https://github.com/rsummers11/CADLab/tree/master/Lung_Segmentation_XLSor).

**Keywords:** Lung segmentation, chest X-ray, criss-cross attention, radiorealistic data augmentation

## 1. Introduction

Lung diseases and disorders are one of the leading causes of death and hospitalization throughout the world. According to the American Lung Association, lung cancer is the number one cancer killer of both women and men in the United States, and more than 33 million Americans are facing a chronic lung disease. The chest radiograph (chest X-ray, or CXR) is one of the most requested radiologic examination for pulmonary diseases such as lung cancer, chronic obstructive pulmonary disease (COPD), pneumonia, tuberculosis, etc. There are huge demands on developing computer-aided diagnosis/detection (CADx/CADe) methods to assist radiologists and other physicians in reading and comprehending chest X-ray images (Shin et al., 2016; Wang et al., 2017, 2018b; Tang et al., 2018c), given the fact that there is a shortage of experienced radiologists, especially in developing

---

\* Contributed equally

countries. Precise segmentation of lung fields can provide rich structural information such as shape irregularity, size measurement and total lung volume, which further facilitates subsequent stages of automated diagnosis (*e.g.*, disease pattern recognition, segmentation and quantization) to assess certain serious clinical conditions.

Over the past decades, automated segmentation of lung boundaries in CXR has received substantial attention in the literature (Candemir et al., 2014; Dai et al., 2017) but still remained a challenging problem (El-Baz et al., 2016). Previous work mainly adopted hand-crafted features to design rule-based systems (Li et al., 2001), active shape/appearance models (Xu et al., 2012), or their hybrid methods (Candemir et al., 2014) to segment the lung boundaries. These approaches rely on the test CXR images being well modeled by the existing training images but they may fail on a different distribution or population. Recently, deep learning based methods (*e.g.* fully convolutional neural networks (FCN) (Shelhamer et al., 2017)) have achieved great successes in biomedical image segmentation (Chen et al., 2018; Tang et al., 2019a; Cai et al., 2018; Tang et al., 2018a) and other medical image analysis tasks (Tang et al., 2019d,c,b, 2018b; Jin et al., 2018; Yan et al., 2018, 2019). The FCN-based methods are intrinsically limited to local receptive fields and insufficient contextual information due to the fixed geometric structures of the convolution. These limitations impose unfavorable effects in segmenting boundaries around less clear lung regions caused by pathological conditions or poor image quality (*e.g.*, low contrast, costophrenic angle clipped off, bad positioning of the patient). Structure correcting adversarial network (SCAN) (Dai et al., 2017) incorporates FCN and adversarial learning (Goodfellow et al., 2014) to segment organs (lungs and heart) in CXRs. SCAN imposes regularization based on the physiological (global) structures by using a critic network that discriminates between the ground truth annotations from the segmentation masks generated by the FCN.

In order to capture richer global contextual information for robust and accurate lung segmentation, we make use of a criss-cross attention (CCA) module (Huang et al., 2018b) to aggregate long-range pixel-wise contextual information in both horizontal and vertical directions. Further dense contextual information can be achieved by stacking more CCA modules recurrently to cover all the pixels. In addition, since publicly available datasets only contain small numbers of lung masks and they are mainly for normal lungs and lungs with subtle findings or unique pathology in an posterioranterior view (*e.g.*, small nodules within the lung field in the JSRT database (Shiraishi et al., 2000), CXRs with tuberculosis presented in the Montgomery database (Jaeger et al., 2014)), it is insufficient to directly use these datasets for training a powerful lung segmentor that can be adapted to pathological lungs with hazy lung boundaries (*e.g.*, large masses, pneumonias, effusions, etc.) for both posterioranterior (PA) and anteroposterior (AP) views. Furthermore, it is very time consuming and tedious for radiologists to manually annotate lung masks, especially on CXRs with abnormalities/pathologies in lung regions (or the so-called abnormal CXRs in this paper). Therefore, we use an image-to-image translation method (Huang et al., 2018a) to synthesize radiorealistic (*i.e.* a synthesized radiograph that appears anatomically realistic) abnormal CXRs from the source of normal ones for data augmentation and mask propagation. The lung masks of synthetic abnormal CXRs are transferred from their normal counterpart and then used as pseudo masks for segmentor retraining.

The proposed framework **XLSor** (*i.e.* **X**-ray **L**ung **S**egmentor) takes advantage of radiorealistic synthesized abnormal CXRs and pseudo masks, without requiring paired normal and abnormal CXRs from the same patient (which is infeasible in reality), as well as the criss-cross attention module to generate robust and accurate lung segmentation. We annotate 100 lung masks on a more

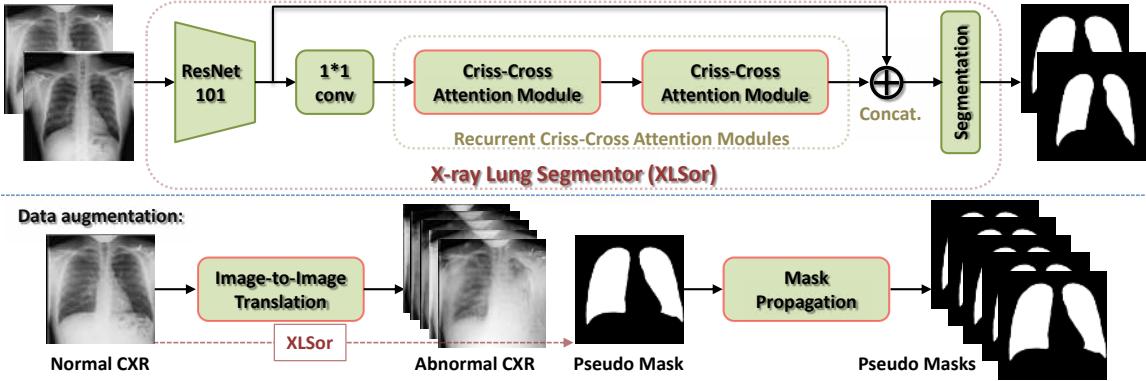


Figure 1: Framework of the proposed X-ray lung segmentor (XLSor).

challenging NIH Chest X-ray dataset (Wang et al., 2017) containing both PA and AP views for evaluation. Extensive experiments on different datasets validate the robustness and effectiveness of the proposed framework.

## 2. Methodology

### 2.1. XLSor Framework Overview

The overall XLSor framework is shown in Figure 1. Given a training set  $R$  with ground-truth masks, an initial lung segmentor is trained (see details in Sec. 2.2). Then, for an auxiliary external set, an image-to-image method MUNIT (Huang et al., 2018a) is used to synthesize abnormal CXRs from normal ones, so as to augment the training data and pseudo mask annotations (mask of normal CXR is obtained using the initial lung segmentor and propagated to its synthesized abnormal CXRs, see details in Sec. 2.3). The initial lung segmentor is updated using  $R$  along with the augmented dataset  $A$  with pseudo masks.

### 2.2. Criss-Cross Attention based Network for Lung Segmentation

In preliminary experiments, we trained a U-Net model (Ronneberger et al., 2015), a widely used model in many applications of medical image segmentation, for lung segmentation. When testing it on the unseen abnormal CXRs, the segmentations are not very promising. That is because the features are extracted from local receptive fields and cannot well capture sufficient contextual information of lungs in U-Net. However, the rich and global contextual information of lungs and their surrounding regions is very important for lung segmentation.

Criss-cross Network (CCNet) (Huang et al., 2018b) achieved state-of-the-art performance in semantic segmentation based on a novel criss-cross attention (CCA) module. Inspired by this, we employ CCA to build a robust and accurate lung segmentor (named XLSor) on chest X-rays. The XLSor is constructed with a fully convolutional network and two CCA modules to capture long-range contextual information (see Figure 1 top). Specifically, we replace the last two down-sampling layers in the ImageNet pre-trained ResNet-101 (He et al., 2016) with dilated convolution operation (Chen et al., 2015), resulting in an output stride of 8. The CCA module collects contextual information in horizontal and vertical directions to enhance pixel-wise representative capability.

Recurrent criss-cross attention module can capture dense contextual information from all pixels by stacking two CCA modules with shared weights. CCA shares the similar idea of capturing global contextual information as the non-local neural network (Wang et al., 2018a) but with much higher computational efficiency. Please refer to (Huang et al., 2018b) for more details about the CCA module. Therefore, the CCA based XLSor can generate clear lung boundaries for more accurate lung segmentation by considering the richer and global contextual information.

The mean square error loss function and the SGD with momentum of 0.9 and weight decay of 0.0005 are used to optimize the XLSor. The initial learning rate is 0.02 and updated using a poly learning rate policy where the initial learning rate is multiplied by  $1 - (\frac{\text{iter}}{\text{max\_iter}})^{0.9}$ , where  $\text{iter}$  is the number of current iterations and  $\text{max\_iter}$  is the total number of iterations. The batch size is set as 4. The size of the input CXR is  $512 \times 512$ .

### 2.3. Data Augmentation via Abnormal Chest X-Ray Pairs Construction

As discussed in Sec. 1, it is insufficient to train a robust lung segmentor using the existing datasets and mask annotations. A simple solution is to enrich the training data, which has been widely used in deep learning. The traditional data augmentation means is to use a combination of affine transformations to manipulate the training data, *e.g.*, shifting, zooming in/out, rotation, flipping, etc, so as to generate new duplicate images for each input image. The contextual information in these generated images do not change very much. To solve these problems, we propose a data augmentation strategy using an image-to-image translation method (Huang et al., 2018a) to construct a large number of abnormal chest X-ray pairs without involving any human intervention, based on which a powerful model can be learned for robust and accurate lung segmentation on different challenging CXRs.

To construct the pairs of abnormal CXR and its corresponding lung masks, there are two straightforward ways. One is to convert the abnormal CXRs into normal ones, and then compute the lung masks which serve as the ground truths for the abnormal CXRs. The other one is to convert the normal CXRs into abnormal ones, and then the lung masks segmented on the normal CXRs are considered as the ground truths of the abnormal ones. Here, we prefer the second way, since the lung regions in real normal CXRs are determined while the ones could be different for various generated normal CXRs in the first way. For the image-to-image translation task, *i.e.* from normal CXRs to abnormal ones, a state-of-the-art method, *i.e.* MUNIT (Huang et al., 2018a), is utilized in this work. MUNIT assumes that the image representation can be decomposed into a content code that is domain-invariant, and a style code that captures domain-specific properties. To translate an image to another domain, MUNIT recombines its content code with a random style code sampled from the style space of the target domain. Please refer to (Huang et al., 2018a) for more details about MUNIT. In this work, we first train the MUNIT model using the default parameter configuration and the NIH chest X-Ray dataset (Wang et al., 2017), from which 5,000 normal CXRs and 5,000 abnormal CXRs are randomly selected for training. Then, given a normal CXR (see Figure 2(a)), we use the trained MUNIT model to generate (or synthesize) a number of abnormal CXRs (see Figure 2(c)-(g)) by combining the content code of the normal CXR and different random style codes learned from the domain of abnormal CXRs. From Figure 2(c)-(g), we can see that the generated abnormal CXRs are radiorealistic. We also notice that the shape of lungs are distorted slightly in the generated abnormal CXRs sometimes. Therefore, the generated abnormalities are customized using the style codes and visually radiorealistic. At last, we use the initial XLSor model trained from the publicly available datasets to obtain the lung masks (see Figure 2(b)) of the given normal

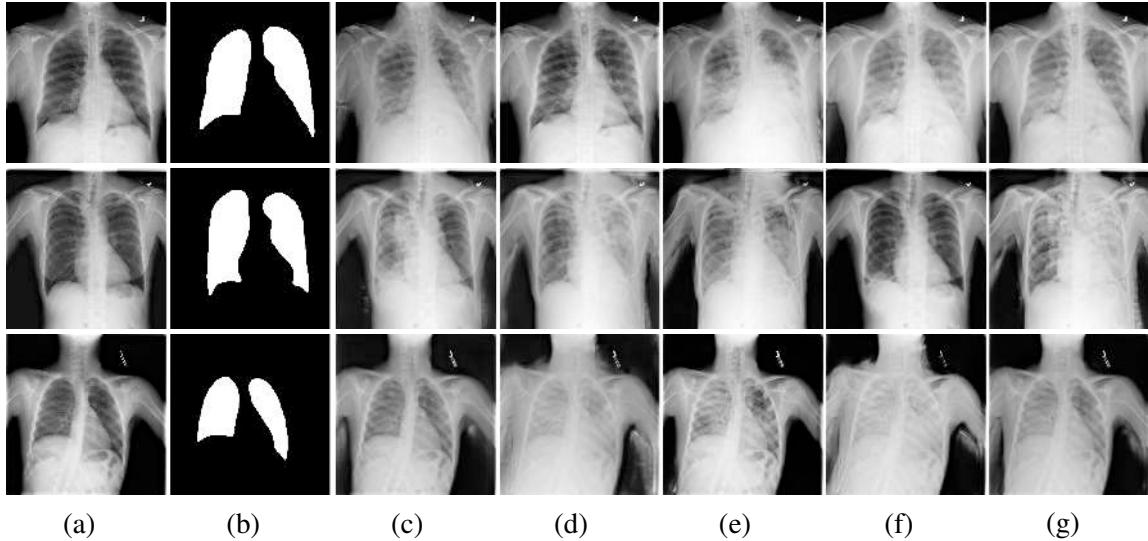


Figure 2: Three examples (rows) of the constructed abnormal CXR pairs. Given an unseen normal CXR (a), XLSor outputs a lung segmentation that is binarized with a threshold of 0.5 to get the lung mask (b) and MUNIT generates different abnormal CXRs (c-g). The lung mask (b) and the synthesized abnormal CXRs (c-g) form the constructed abnormal CXR pairs.

CXR, which are also considered as the pseudo masks of the generated abnormal CXRs (*i.e.* mask propagation) to form the constructed abnormal CXR pairs (see Figure 1 bottom) for further training the XLSor model. We also iteratively conducted above processes and found that it is not helpful because the normal CXRs are easy to segment and the pseudo masks are good enough at the first iteration.

### 3. Experiments

#### 3.1. Datasets and Evaluation Criteria

We evaluate the lung segmentation performance of the proposed XLSor using two publicly available datasets, *i.e.* JSRT (Shiraishi et al., 2000) and Montgomery (Jaeger et al., 2014), and our own annotated dataset (named NIH). **JSRT** contains 247 CXRs, among which 154 have lung nodules and 93 have no lung nodule. **Montgomery** contains 138 CXRs, including 80 normal patients and 58 patients with manifested tuberculosis (TB). Both datasets provide pixel-wise lung mask annotations. We notice that the abnormal lung regions in these two datasets are mild. Only using such datasets for evaluation cannot well demonstrate the effectiveness and generalizability of the methods, since diseases can occasionally cause severe damages to the lungs. Therefore, we manually annotate the lung masks of 100 abnormal CXRs with various severity of lung diseases, which are selected from the NIH Chest X-Ray dataset (Wang et al., 2017) by excluding the samples used for MUNIT training. Here, we name the manually labeled set as **NIH**.

JSRT and Montgomery datasets are combined and randomly split into three subsets for both normal and abnormal CXRs, *i.e.* training (70%), validation (10%) and testing (20%). Specifically, the validation and testing sets include 37 and 78 CXRs, respectively. The remaining 280 CXRs serve as a training set for model training. The validation set is used for model selection, and the testing set and the NIH dataset are used for performance evaluation. Five criteria, *i.e.* volumetric similarity (VS), averaged Hausdorff distance (AVD), Dice similarity coefficient (DICE), precision (PRE) and recall (REC) scores, are calculated pixel-wisely by a publicly available segmentation evaluation tool ([Taha and Hanbury, 2015](#)) with threshold of 0.5 and used to evaluate the quantitative segmentation performance.

### 3.2. Quantitative Results

In this work, U-Net ([Ronneberger et al., 2015](#)) is applied for performance comparisons to demonstrate the effectiveness of the criss-cross attention based XLSor. To validate the usefulness of adding the augmented samples for lung segmentation, we first use the proposed data augmentation strategy to generate four augmented training sets, denoted as  $A^1$ ,  $A^2$ ,  $A^3$  and  $A^4$ , respectively. Here,  $A^1$  contains 600 constructed pairs including 100 normal pairs and 500 abnormal pairs where five abnormal CXRs are synthesized from each normal CXR using MUNIT ([Huang et al., 2018a](#)).  $A^i$  ( $i = 2, 3, 4$ ) contains all samples in  $A^{i-1}$  and another new 600 constructed pairs. We then train the XLSor and U-Net models for lung segmentation using six different training settings, *i.e.* only using the real public training set (denoted  $R$ ), using the real public training set and any augmented set  $A^i$  ( $i = 1, 2, 3, 4$ ) (denoted  $R + A^i$ ), and only using the augmented set  $A^4$ . To validate the effectiveness of CCA for segmentation performance improvement, we also train the XLSor model without CCA modules (denoted  $XLSor^-$ ) and the U-Net model with CCA modules (denoted  $U\text{-Net}^+$ ) using  $R$  and  $R + A^4$ . In each training setting, the same traditional data augmentation techniques (*e.g.*, scaling and flipping) are adopted. Finally, the five criteria are used to evaluate the performance of lung segmentation on the public testing set and NIH dataset, whose results are reported in Table 1.

From Table 1, we can see that 1) the proposed XLSor gets better results than U-Net on both the simple public testing set and the difficult NIH dataset. Especially, the performance of  $XLSor_R$  is much better than the one of  $U\text{-Net}_R$  on the NIH dataset (*e.g.*, improving the Dice score about 12%), meaning that the proposed XLSor is able to work much better than U-Net on the unseen CXRs whose data distribution is much different from the training data. This demonstrates that the proposed XLSor based on the criss-cross attention module can well learn the global contextual information of lung regions and strong discriminative features to distinguish the lung regions from their surrounding structures regardless of the CXRs' properties. 2) When adding the augmented samples for model training, the performance is improved, *i.e.*  $XLSor_{R+A^i}$  (or  $U\text{-Net}_{R+A^i}$ ) gets better results than  $XLSor_R$  (or  $U\text{-Net}_R$ ), suggesting the effectiveness of our data augmentation technique for lung segmentation performance improvement. Through experiments, we find that the performance remains stable when adding more augmented samples than  $A^4$ . 3) When only using the augmented samples for model training, both XLSor and U-Net still get very promising performance on the public testing set and the NIH dataset (see the results of  $XLSor_{A^4}$  and  $U\text{-Net}_{A^4}$  in Table 1), suggesting that the generated abnormal CXRs are radiorealistic and the pseudo lung masks effectively supervise the learning processes for lung segmentation. 4) The results by all models are quite similar in the public testing set, that is because the testing CXRs are all (near-)normal and the lung segmentation task is relatively easy. 5) U-Net obtains worse performance on NIH dataset than the

Table 1: Lung segmentation results on the **public testing set** and **NIH dataset** using the proposed XLSor and U-Net with different training settings. Results showing mean with standard deviation.  $\uparrow$ : the larger the better.  $\downarrow$ : the smaller the better.

Method	REC $\uparrow$	PRE $\uparrow$	DICE $\uparrow$	AVD $\downarrow$	VS $\uparrow$
<i>Public testing set</i>					
XLSor <sub>R</sub>	0.973 $\pm$ 0.02	<b>0.979<math>\pm</math>0.02</b>	<b>0.976<math>\pm</math>0.01</b>	0.149 $\pm$ 0.51	<b>0.992<math>\pm</math>0.01</b>
XLSor <sub>R+A<sup>1</sup></sub>	0.973 $\pm$ 0.02	<b>0.979<math>\pm</math>0.02</b>	<b>0.976<math>\pm</math>0.01</b>	0.152 $\pm$ 0.52	0.991 $\pm$ 0.01
XLSor <sub>R+A<sup>2</sup></sub>	<b>0.974<math>\pm</math>0.02</b>	0.978 $\pm$ 0.02	<b>0.976<math>\pm</math>0.01</b>	<b>0.117<math>\pm</math>0.31</b>	0.991 $\pm$ 0.01
XLSor <sub>R+A<sup>3</sup></sub>	0.972 $\pm$ 0.02	<b>0.979<math>\pm</math>0.02</b>	<b>0.976<math>\pm</math>0.01</b>	0.126 $\pm$ 0.33	0.991 $\pm$ 0.01
XLSor <sub>R+A<sup>4</sup></sub>	<b>0.974<math>\pm</math>0.02</b>	0.977 $\pm$ 0.02	<b>0.976<math>\pm</math>0.01</b>	0.146 $\pm$ 0.44	0.991 $\pm$ 0.01
XLSor <sub>A<sup>4</sup></sub>	0.965 $\pm$ 0.03	<b>0.979<math>\pm</math>0.02</b>	0.972 $\pm$ 0.02	0.162 $\pm$ 0.36	0.989 $\pm$ 0.01
XLSor <sub>R</sub> <sup>-</sup>	0.973 $\pm$ 0.02	0.978 $\pm$ 0.02	0.975 $\pm$ 0.01	0.151 $\pm$ 0.53	0.991 $\pm$ 0.01
XLSor <sub>R+A<sup>4</sup></sub> <sup>-</sup>	0.972 $\pm$ 0.02	0.978 $\pm$ 0.02	0.976 $\pm$ 0.01	0.148 $\pm$ 0.47	0.991 $\pm$ 0.01
U-Net <sub>R</sub>	0.976 $\pm$ 0.02	0.968 $\pm$ 0.03	0.972 $\pm$ 0.02	0.198 $\pm$ 0.56	0.988 $\pm$ 0.02
U-Net <sub>R+A<sup>1</sup></sub>	0.973 $\pm$ 0.02	0.976 $\pm$ 0.02	0.974 $\pm$ 0.01	0.162 $\pm$ 0.54	<b>0.990<math>\pm</math>0.01</b>
U-Net <sub>R+A<sup>2</sup></sub>	<b>0.977<math>\pm</math>0.02</b>	0.973 $\pm$ 0.02	<b>0.975<math>\pm</math>0.01</b>	0.135 $\pm$ 0.41	0.989 $\pm$ 0.01
U-Net <sub>R+A<sup>3</sup></sub>	0.976 $\pm$ 0.02	0.975 $\pm$ 0.02	<b>0.975<math>\pm</math>0.01</b>	<b>0.131<math>\pm</math>0.34</b>	<b>0.990<math>\pm</math>0.01</b>
U-Net <sub>R+A<sup>4</sup></sub>	0.973 $\pm$ 0.02	<b>0.978<math>\pm</math>0.01</b>	<b>0.975<math>\pm</math>0.01</b>	0.152 $\pm$ 0.46	<b>0.990<math>\pm</math>0.01</b>
U-Net <sub>A<sup>4</sup></sub>	0.967 $\pm$ 0.02	0.975 $\pm$ 0.01	0.971 $\pm$ 0.01	0.164 $\pm$ 0.37	0.989 $\pm$ 0.01
U-Net <sub>R</sub> <sup>+</sup>	0.976 $\pm$ 0.02	0.970 $\pm$ 0.03	0.972 $\pm$ 0.02	0.191 $\pm$ 0.54	0.988 $\pm$ 0.02
U-Net <sub>R+A<sup>4</sup></sub> <sup>+</sup>	0.975 $\pm$ 0.02	0.977 $\pm$ 0.01	0.975 $\pm$ 0.01	0.130 $\pm$ 0.33	0.990 $\pm$ 0.01
<i>NIH dataset</i>					
XLSor <sub>R</sub>	0.966 $\pm$ 0.02	0.927 $\pm$ 0.09	0.943 $\pm$ 0.05	0.669 $\pm$ 1.64	0.966 $\pm$ 0.05
XLSor <sub>R+A<sup>1</sup></sub>	0.958 $\pm$ 0.03	0.973 $\pm$ 0.02	0.965 $\pm$ 0.02	0.172 $\pm$ 0.26	0.985 $\pm$ 0.01
XLSor <sub>R+A<sup>2</sup></sub>	0.962 $\pm$ 0.02	0.980 $\pm$ 0.01	0.971 $\pm$ 0.01	0.097 $\pm$ 0.08	0.989 $\pm$ 0.01
XLSor <sub>R+A<sup>3</sup></sub>	0.967 $\pm$ 0.02	0.978 $\pm$ 0.02	0.973 $\pm$ 0.01	0.089 $\pm$ 0.07	0.990 $\pm$ 0.01
XLSor <sub>R+A<sup>4</sup></sub>	<b>0.974<math>\pm</math>0.01</b>	0.976 $\pm$ 0.01	<b>0.975<math>\pm</math>0.01</b>	<b>0.078<math>\pm</math>0.06</b>	<b>0.993<math>\pm</math>0.01</b>
XLSor <sub>A<sup>4</sup></sub>	0.964 $\pm$ 0.02	<b>0.983<math>\pm</math>0.01</b>	0.973 $\pm$ 0.01	0.098 $\pm$ 0.13	0.988 $\pm$ 0.01
XLSor <sub>R</sub> <sup>-</sup>	0.965 $\pm$ 0.03	0.902 $\pm$ 0.10	0.929 $\pm$ 0.06	0.952 $\pm$ 1.81	0.955 $\pm$ 0.06
XLSor <sub>R+A<sup>4</sup></sub> <sup>-</sup>	0.965 $\pm$ 0.02	0.981 $\pm$ 0.01	0.967 $\pm$ 0.01	0.093 $\pm$ 0.10	0.990 $\pm$ 0.01
U-Net <sub>R</sub>	0.938 $\pm$ 0.07	0.761 $\pm$ 0.20	0.823 $\pm$ 0.16	5.231 $\pm$ 9.02	0.869 $\pm$ 0.15
U-Net <sub>R+A<sup>1</sup></sub>	0.926 $\pm$ 0.05	<b>0.960<math>\pm</math>0.03</b>	0.942 $\pm$ 0.03	0.832 $\pm$ 1.29	0.971 $\pm$ 0.02
U-Net <sub>R+A<sup>2</sup></sub>	0.947 $\pm$ 0.04	0.950 $\pm$ 0.04	0.948 $\pm$ 0.03	0.500 $\pm$ 1.03	0.981 $\pm$ 0.02
U-Net <sub>R+A<sup>3</sup></sub>	0.950 $\pm$ 0.03	0.954 $\pm$ 0.03	0.951 $\pm$ 0.02	0.393 $\pm$ 0.58	0.983 $\pm$ 0.02
U-Net <sub>R+A<sup>4</sup></sub>	0.943 $\pm$ 0.04	0.958 $\pm$ 0.03	0.950 $\pm$ 0.03	0.454 $\pm$ 0.73	0.982 $\pm$ 0.02
U-Net <sub>A<sup>4</sup></sub>	<b>0.952<math>\pm</math>0.03</b>	0.959 $\pm$ 0.03	<b>0.955<math>\pm</math>0.02</b>	<b>0.315<math>\pm</math>0.47</b>	<b>0.983<math>\pm</math>0.02</b>
U-Net <sub>R</sub> <sup>+</sup>	0.929 $\pm$ 0.07	0.804 $\pm$ 0.20	0.842 $\pm$ 0.14	4.782 $\pm$ 8.05	0.895 $\pm$ 0.14
U-Net <sub>R+A<sup>4</sup></sub> <sup>+</sup>	0.956 $\pm$ 0.03	0.969 $\pm$ 0.02	0.962 $\pm$ 0.02	0.262 $\pm$ 0.54	0.985 $\pm$ 0.02

public testing set, meaning that the CXRs in the NIH dataset are more complex and difficult than the ones in the public testing set. But XLSor can get comparable and good results on both datasets, sug-

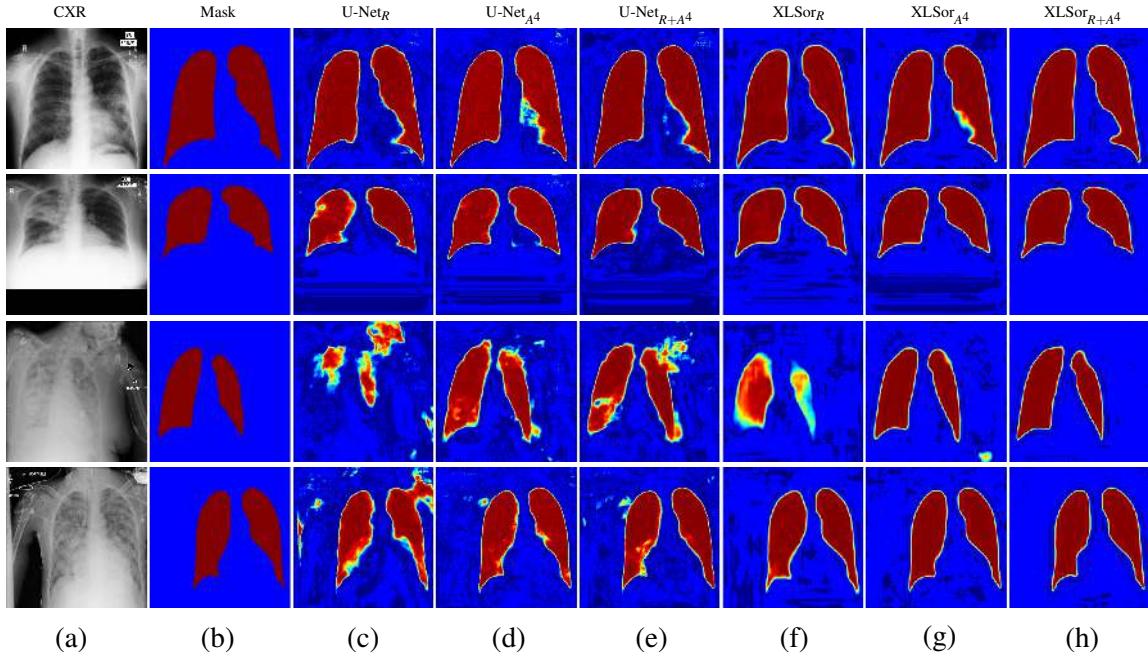


Figure 3: Four examples (rows) of lung segmentation results produced by XLSor and U-Net trained using  $R$ ,  $A^4$  and  $R + A^4$ . Here, the results are given as the probability maps directly outputted by the models, which can be binarized with a threshold of 0.5 to get the binary lung masks for performance evaluation. The first two rows are from the public testing set and the last two rows are from the NIH dataset. To better visualize the differences between lung segmentation results and ground truths, we colorize them with pseudo-colors. Better viewed in color.

gesting that the proposed XLSor is robust and powerful for lung segmentation in different scenarios. 6) XLSor/U-Net<sup>+</sup> achieves better results than XLSor<sup>-</sup>/U-Net (especially, on the NIH dataset), suggesting that using CCA modules can make the model learn the global contextual information of lung regions better and extract more powerful discriminative features for performance improvement. All results quantitatively demonstrate the effectiveness and generalizability of the proposed XLSor for lung segmentation on various CXRs.

### 3.3. Qualitative Results

Figure 3 shows four qualitative lung segmentation results produced by the models (*i.e.* XLSor and U-Net) trained with the following settings:  $R$ ,  $A^4$  and  $R + A^4$ . Compared with U-Net, the lung segmentation results produced by the proposed XLSor are much closer to the ground truths in various challenging scenarios. To be specific, 1) the proposed XLSor not only highlights the correct lung regions clearly, but also well suppresses the probabilities of background regions, so as to produce the segmentation results with higher contrast between lung regions and background than U-Net. 2) With the help of the criss-cross attention module that considers sufficient contextual information, the proposed XLSor is able to output the lung segmentations with clear boundaries

and consistent probabilities, even when the model is trained and tested on CXRs with different distribution of abnormalities. 3) With the augmented samples for training, the qualities of lung segmentations are improved. These intuitively demonstrate the effectiveness of the proposed XLSor and the usefulness of the proposed data augmentation strategy for lung segmentation on chest X-rays.

#### 4. Conclusions and Future Work

In this paper, we propose a robust and accurate lung segmentor based on a criss-cross attention network and a customized radiorealistic abnormalities generation technique for data augmentation. Experiments showed that the proposed framework was able to capture rich contextual information from both original and radiorealistic synthesized CXRs to adapt to more challenging images, resulting in much better segmentation, especially in unseen abnormal CXRs. Future work includes segmenting more organs and integrating with more downstream tasks such as disease classification and detection to provide comprehensive and accurate computer-aided detection on CXR images, *e.g.*, performing segmentation and classification simultaneously by training different MUNIT models for individual diseases and using them to generate abnormalities accordingly in categories.

#### Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health Clinical Center and by the Ping An Insurance Company through a Cooperative Research and Development Agreement. We thank Nvidia for GPU card donation.

#### References

- J. Cai, Y. Tang, L. Lu, A. P Harrison, K. Yan, J. Xiao, L. Yang, and R. M. Summers. Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3D mask generation from 2D RECIST. In *MICCAI*, 2018.
- S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE TMI*, 33(2):577–590, 2014.
- L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert. Drinet for medical image segmentation. *IEEE TMI*, 37(11):2453–2462, 2018.
- L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015.
- W. Dai, J. Doyle, X. Liang, H. Zhang, N. Dong, Y. Li, and E. P. Xing. Scan: Structure correcting adversarial network for chest x-rays organ segmentation. *arXiv preprint arXiv:1703.08770*, 2017.
- A. El-Baz, X. Jiang, and J.S. Suri. *Biomedical Image Segmentation: Advances and Trends*. CRC Press, Taylor & Francis Group, 2016. ISBN 9781482258554.
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- X. Huang, M.Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018a.
- Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. CCNet: Criss-cross attention for semantic segmentation. *arXiv preprint arXiv:1811.11721*, 2018b.
- S. Jaeger, S. Candemir, Y. X. Antani, S. and Wang, P. X. Lu, and G. Thoma. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *QIMS*, 4(6):475–477, 2014.
- D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura. CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. In *MICCAI*, 2018.
- L. Li, Y. Zheng, M. Kallergi, and R. A. Clark. Improved method for automatic identification of lung regions on chest radiographs. *AR*, 8(7):629 – 638, 2001.
- O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.
- E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE TPAMI*, 39(4):640–651, 2017.
- H. C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers. Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation. In *CVPR*, 2016.
- J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi. Development of a digital image database for chest radiographs with and without a lung nodule. *AJR*, 174(1):71–74, 2000.
- A. A. Taha and A. Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC MI*, 15(1):29, 2015.
- Y. Tang, J. Cai, L. Lu, A. P. Harrison, K. Yan, J. Xiao, L. Yang, and R. M. Summers. CT image enhancement using stacked generative adversarial networks and transfer learning for lesion segmentation improvement. In *MLMI*, 2018a.
- Y. Tang, A. P. Harrison, M. Bagheri, J. Xiao, and R. M. Summers. Semi-automatic RECIST labeling on CT scans with cascaded convolutional neural networks. In *MICCAI*, 2018b.
- Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In *MLMI*, 2018c.
- Y. B. Tang, S. Oh, Y. X. Tang, J. Xiao, and R. M. Summers. CT-realistic data augmentation using generative adversarial network for robust lymph node segmentation. In *Medical Imaging: CAD*, 2019a.
- Y. B. Tang, K. Yan, Y. X. Tang, J. Liu, J. Xiao, and R. M. Summers. ULDor: A universal lesion detector for CT scans with pseudo masks and hard negative example mining. In *ISBI*, 2019b.

- Y. X. Tang, Y. B. Tang, M. Han, J. Xiao, and R. M. Summers. Abnormal chest X-ray identification with generative adversarial one-class classifier. In *ISBI*, 2019c.
- Y. X. Tang, Y. B. Tang, M. Han, J. Xiao, and R. M. Summers. Deep adversarial one-class learning for normal and abnormal chest radiograph classification. In *Medical Imaging: CAD*, 2019d.
- X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017.
- X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018a.
- X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In *CVPR*, 2018b.
- T. Xu, M. Mandal, R. Long, I. Cheng, and A. Basu. An edge-region force guided active shape approach for automatic lung field detection in chest radiographs. *CMIG*, 36(6):452 – 463, 2012.
- K. Yan, X. Wang, L. Lu, L. Zhang, A. P. Harrison, M. Bagheri, and R. M. Summers. Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In *CVPR*, 2018.
- K. Yan, Y. Peng, Z. Lu, and R. M. Summers. Fine-grained lesion annotation in CT images with knowledge mined from radiology reports. In *CVPR*, 2019.

# Training Deep Networks on Domain Randomized Synthetic X-ray Data for Cardiac Interventions

**Daniel Toth<sup>1,2</sup>**

DANIEL.TOTH@KCL.AC.UK

<sup>1</sup> Siemens Healthineers, Frimley, UK

<sup>2</sup> School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK

**Serkan Cimen<sup>3</sup>**

**Pascal Ceccaldi<sup>3</sup>**

<sup>3</sup> Siemens Healthineers, Medical Imaging Technologies, Princeton, NJ, USA

**Tanja Kurzendorfer<sup>4</sup>**

<sup>4</sup> Siemens Healthineers, Forchheim, Germany

**Kawal Rhode<sup>\*2</sup>**

**Peter Mountney<sup>\*3</sup>**

## Abstract

One of the most significant challenges of using machine learning to create practical clinical applications in medical imaging is the limited availability of training data and accurate annotations. This problem is acute in novel multi-modal image registration applications where complete datasets may not be collected in standard clinical practice, data may be collected at different times and deformation makes perfect annotations impossible. Training machine learning systems on fully synthetic data is becoming increasingly common in the research community. However, transferring to real world applications without compromising performance is highly challenging. Transfer learning methods adapt the training data, learned features, or the trained models to provide higher performance on the target domain. These methods are designed with the available samples, but if the samples used are not representative of the target domain, the method will overfit to the samples and will not generalize. This problem is exacerbated in medical imaging, where data of the target domain is extremely scarce. This paper proposes to use Domain Randomization (DR) to bridge the reality gap between the training and target domains, requiring no samples of the target domain. DR adds unrealistic perturbations to the training data, such that the target domain becomes just another variation. The effects of DR are demonstrated on a challenging task: 3D/2D cardiac model-to-X-ray registration, trained fully on synthetic data generated from 1711 clinical CT volumes. A thorough qualitative and quantitative evaluation of transfer to clinical data is performed. Results show that without DR training parameters have little influence on performance on the training domain of digitally reconstructed radiographs, but can cause substantial variation on the target domain (X-rays). DR results in a significantly more consistent transfer to the target domain.

**Keywords:** Domain Randomization, Imitation Learning, Cardiac Registration.

## 1. Introduction

In recent years, deep learning algorithms have been successfully applied to a wide variety of medical imaging tasks ([Litjens et al., 2017](#)). However, for many medical imaging problems, labeled

---

\* Joint senior authors

training data is extremely scarce. Especially in novel clinical procedures, new, procedure-specific, retrospectively not available data might be required. Initially, data is available only in highly limited quantities and it can accumulate very slowly to usable amounts. Multimodal datasets, e.g., for image registration, can be even more challenging to acquire and previously collected data is highly unlikely to be multimodal. Additionally, Ground Truth (GT) annotations can be highly challenging to acquire. The annotation needs to be performed by experts in the respective domain, i.e., by clinicians. This can be time consuming and expensive for even smaller amounts of data, but might not be feasible for larger datasets. In certain applications, it might be impossible to acquire GT annotations, e.g., due to deformations in multimodal data acquired at different time points.

A system for registering 3D preoperative data to 2D X-rays (Toth et al., 2018) demonstrated the feasibility and benefits of training purely on synthetic data. The system generated synthetic images from a CT dataset, thus knowing the GT transformations. These synthetic X-ray images are commonly referred to as Digitally Reconstructed Radiographs (DRRs) (Russakoff et al., 2005). The registration system trained on DRR data has shown good accuracy on the same domain and acceptable performance on clinical X-ray data. However, transfer performance is highly affected by training parameters: different local minima might be reached, resulting in highly similar performance on the synthetic training domain, but highly variable results on the target domain. DRR generation algorithms do not model every aspect of the X-ray formation. This causes images to appear differently than clinically acquired X-rays. Algorithms used for image guidance have to be extremely robust against perturbations. In the case of the X-ray data, these could be intensity variations, due to different dose settings, variation of the collimation, or devices in the field of view. These variations can appear differently, or may not be present, in the artificially generated DRRs. The transfer, thus the robustness of the registration can suffer. The performance reduction after transferring to the target domain is called the reality gap.

Several approaches have addressed this issue including (Heimann et al., 2014), where a classifier is adapted to the target domain, without requiring labelled data on the target domain. However, it could fail, if the training and target distributions are far. Domain adaptation can also be performed through task-driven Generative Adversarial Networks (GANs) (Zhang et al., 2018). GANs, however, can be difficult to train, can be unstable, can show mode collapse. Advanced simulation methods can be used to generate higher quality DRRs that appear more similar to the target domain's images, such as (Unberath et al., 2018). Such an approach is still not a perfect X-ray simulator, important features might not be simulated. The collective weakness of these approaches is that they try to adapt/fit to samples of the target domain. If the samples are not representative of the target domain, the adaptation will overfit to the samples. If such a system has to process data from a new site, device, or of unexpected imaging parameters, performance can decrease significantly.

This paper addresses the problem of overfitting to the target domain, by using Domain Randomization (DR) (Tobin et al., 2017). In DR the training data is augmented with unrealistic transformations. The transformations introduce such a large variation to the data that the target domain will appear as just another variation to the learning system. This approach was shown to help transfer to the target domain in non medical applications, e.g., autonomous driving (Tremblay et al., 2018) or robotic control (Peng et al., 2018).

In this paper, the usability of DR in the medical imaging context is demonstrated for the first time. Its benefits are being shown on the challenging cardiac image registration task described in (Toth et al., 2018). It is shown that, if trained only on artificially generated data (DRRs), DR can greatly increase robustness against geometric perturbations. Furthermore, it is demonstrated that,

without DR, variation of training parameters can result in inconsistent results on the target domain. Performance with DR, however, appears to be agnostic to variation in training parameters, thus produces more consistent results.

## 2. Methods

The registration framework with domain randomization is illustrated in Figure 1. The model-to-X-ray registration system from (Toth et al., 2018) is intended to be used for 3D/2D cardiac registration. It is trained purely with CT volumes, without any real X-rays from the target domain. The CTs are used for: 1) DR DRRs and 2) to segment, perturb, and project masks of the Left Ventricle (LV) into 2D. Since both images are generated, the GT transformation is known between them. This is used to compute rewards for each possible action of the three degrees of freedom given by the imaging plane: translations  $x$ ,  $y$  and rotation about the  $z$  axis. The rewards and the two 2D images (DR DRR and mask image) are used to train an artificial agent represented by a Convolutional Neural Network (CNN), see Figure 1a. For inference, the artificial agent is being shown the two images and is able to predict the rewards for each possible action. The action with the highest reward is chosen and is applied to the 3D model, as depicted in Figure 1b.

### 2.1. Transfer Learning

The training of the artificial agent is performed on fully synthetic data. It has shown acceptable performance in terms of accuracy on the synthetic domain and robustness on the target domain. However, for interventional applications, highest levels of robustness are crucial. How well the imitation learning agent transfers to real data is influenced by three main factors: 1) the appearance of the training data (similarity to target data), 2) the weight initialization of the neural network, and 3) the training data order.

The similarity between the training domain and the target domain is the main factor in performance after transfer. If the training DRRs are sufficiently similar to the target X-ray images, the network will transfer with similar performance. This can be achieved to a certain degree by optimizing DRR generation parameters, but it is highly challenging and the learned network might not generalize well to unseen images.

The second factor is the initialization of the weights in the neural network representing the agent. The initial weights define the starting position in the optimization space of training. From

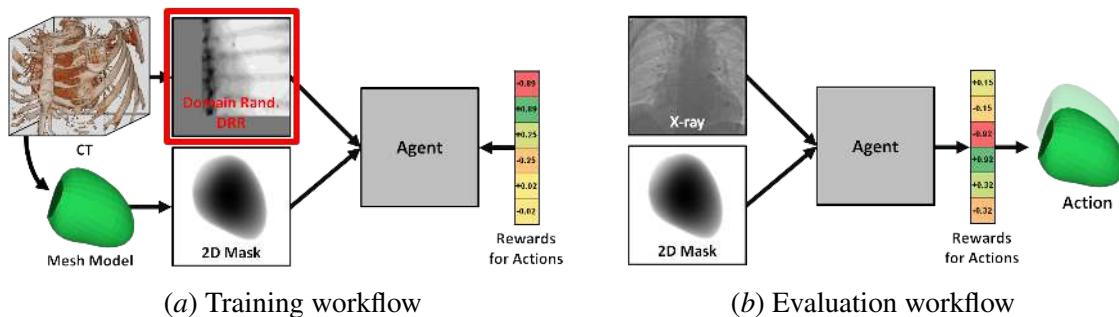


Figure 1: Overview of registration framework workflows.

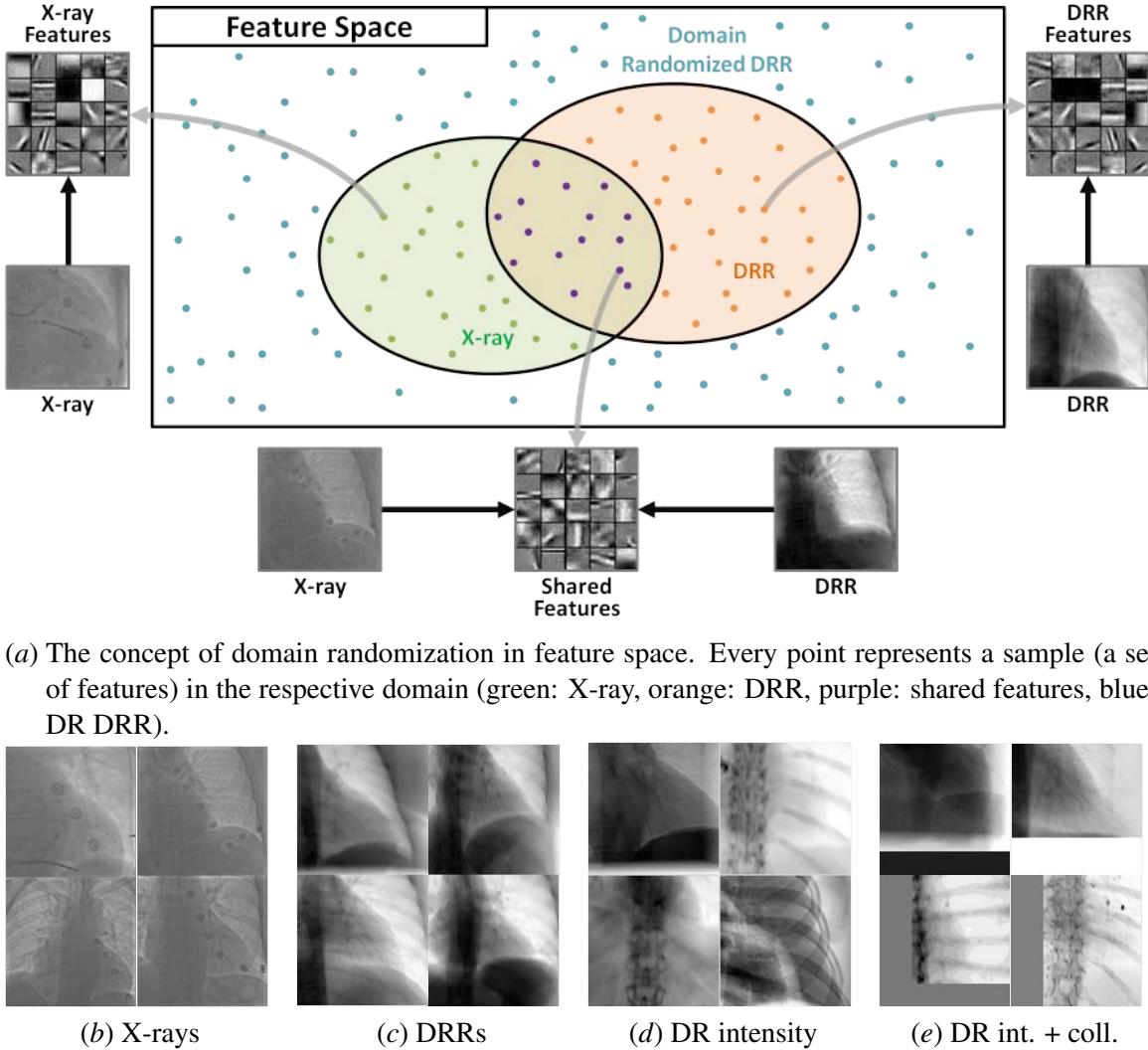


Figure 2: The concept of domain randomization in feature space and sample images.

different starting positions, different local minima can be reached that can result in highly similar results on the training, but inconsistent results on the target domain.

The effects of the variation of the training data order can have similar influence. If the same weights are used for initialization, but the network sees the training data in a different order, the optimization can take different steps and can end up in different local minima, resulting in different results on the target domain.

## 2.2. Domain Randomization

Domain randomization (Tobin et al., 2017) is an approach to transfer a deep neural network trained on only simulated images to the real world. Instead of high quality simulation images, domain randomization relies on the variability of the low quality simulation images. The main idea is to

randomize the simulator and generate a large variation of simulated images, such that the real images become just another variation.

For the present problem, instead of using a realistic X-ray simulator, it is proposed to use a simple ray tracing-based DRR renderer. The concept can be illustrated in feature space by a Venn diagram, see Figure 2a. The sets of X-ray features and DRR features have a substantial intersection, because the method transfers (with limited performance) to the X-ray domain, if trained only on DRRs. If there were no shared features, no intersection of the sets, the approach would not transfer at all. To ensure a more successful transfer, the DRR generation is randomized that the resulting features cover a large portion of the whole feature space (blue in Figure 2a). If the DR images cover most of the feature space, the approach will generalize to the X-rays too, even if no X-ray similar images are generated.

The DRR generation is performed with randomized 1) Hounsfield Unit (HU) values of the 3D CT data and 2) collimation added to the 2D projections. Before computing the ray tracing, an intensity mapping to the HU values of the CT data was applied either globally or locally. For global randomization, a non-linear function (cumulative distribution function of a beta distribution) was applied to transform the voxel values, where the parameters  $\alpha$  and  $\beta$  are randomly sampled from a uniform distribution of [0.5; 5.0]. For the local randomization, the intensity range of the CT data was subdivided into 5 to 15 non-overlapping ranges. The start and end of the intervals was defined by evenly spacing the whole range, then a random perturbation was added to the start/end values sampled from a Gaussian distribution (mean = 80 and standard deviation = 40). For each intensity range, one of three randomization options was chosen: 1) the non-linear mapping described above; 2) invert the intensity range; 3) shift the intensity values by adding a random value of a uniform distribution on the interval [-100; 100]. This results in images shown in Figure 2d.

The acquired X-ray images of the target domain are often collimated. Thus, in addition to the intensity randomization, bands of random intensity were added to the generated images, see Figure 2e. The intensity was varied in the range of possible intensity values and the size of the borders was maximally 25 % of the image size in the respective direction.

### 3. Experiments and Results

The effects of domain randomization are demonstrated on the image registration task with an imitation learning agent, described in Section 2. The agent was trained only on synthetic data generated from 1711 clinical CT volumes of 799 patients. The evaluation was performed in two steps, on two datasets: 1) synthetic data from the same domain as the training data originates from; 2) clinical patient data, from the target domain.

The effect of DR is presented for a certain set of seed values for 1) random weight initialization and 2) random data shuffling. Then the seeds are varied to demonstrate the inconsistency of the transfer from the synthetic training domain to the real target domain and to show the consistency of the domain randomization results. An ablation study was also performed to investigate the influence of each DR.

#### 3.1. Evaluation Data

The synthetic test dataset consists of 100 CT volumes of 100 patients. There is no overlap between training and test patients. The 3D mesh model of each patient was perturbed 10 times sampled from a uniform distribution, resulting in 1000 artificial test cases. Due to the known perturbation, the GT

registration is available. Accuracy is measured by the points of a 3D landmark at the center of the LV model.

The clinical data consists of 21 patient datasets. Each patient dataset has one MR volume of that the mesh model was extracted and an X-ray image to register to. The mesh model is perturbed 169 times in a rectangular grid manner about an approximate, manual registration. There is no GT registration available for this dataset. The registration is evaluated quantitatively for robustness, i.e., that it provides the same result from different perturbations. The robustness error measure was defined as described in (Toth et al., 2018):

$$e_f = \|\mathbf{x}_f - \tilde{\mathbf{x}}_f\|_2, \quad (1)$$

where  $\tilde{\mathbf{x}}_f$  is the median final position of the center of the cross landmark of all perturbations and  $\mathbf{x}_f$  is the final position for a single perturbation. Rotational robustness  $\varepsilon_f$  was computed in the same manner. Accuracy is only verified qualitatively, by comparing the heart shadow in the X-rays and the edges of the projected LV mask.

Table 1: Accuracy in mm on synthetic DRR data of the standard (SN) and domain randomized networks (DRN) for varying weight initialization  $W_i$  and data order  $D_i$ .

(a) Weight initialization varied							(b) Data order varied						
	SN/DRN							SN/DRN					
Start	$W_1, D_1$	$W_2, D_1$	$W_3, D_1$	$W_4, D_1$	$W_5, D_1$		Start	$W_1, D_1$	$W_1, D_2$	$W_1, D_3$	$W_1, D_4$	$W_1, D_5$	
<b>Mean</b>	22.41	2.80/2.65	2.66/2.66	2.70/2.65	2.63/2.64	2.77/2.68	<b>Mean</b>	22.41	2.80/2.65	2.78/2.58	2.64/2.59	2.80/2.78	2.70/2.62
<b>StD.</b>	10.60	2.17/2.05	2.27/2.12	2.18/2.10	2.23/2.07	2.24/2.19	<b>StD.</b>	10.60	2.17/2.05	2.19/1.98	2.17/2.05	2.21/2.23	2.19/2.07
<b>Median</b>	21.27	2.15/2.16	1.99/2.14	2.15/2.14	1.96/2.08	2.23/2.14	<b>Median</b>	21.27	2.15/2.16	2.26/2.03	2.02/2.06	2.25/2.19	2.08/2.10
<b>90 %</b>	37.12	5.54/5.00	5.50/5.34	5.31/5.33	5.26/5.07	5.52/5.24	<b>90 %</b>	37.12	5.54/5.00	5.31/5.11	5.21/5.23	5.39/5.50	5.14/5.05

### 3.2. Baseline

The Standard Network (SN) was trained with a selection of seeds for data shuffling ( $D_1$ ) and weight initialization ( $W_1$ ). The training curves are marginally closer to each other for DR, see Appendix A. The synthetic results are slightly better, see Table 1, but the clinical data results are slightly worse than in (Toth et al., 2018), see Figure 3a. This is due to training on significantly more data, causing the network to overfit more to the synthetic domain. The network with domain randomization (DRN) has the same architecture and was trained with the same parameters. On synthetic data, as expected, the results improve only slightly, because no transfer is required. The clinical data experiment has shown, however, that the distributions are compressed and the number of outliers is greatly reduced, see Figures 3a and 3c. Most individual cases have improved greatly in the final alignment, for samples see Figures 3b and 3d. The network has learned to handle intensity and collimation variations, although, it has never seen realistic transformations.

### 3.3. Parameter Variation

Two experiments were performed: 1) the seed for random network weight initialization was varied, with a fixed data order ( $W_i, D_1$ ) and 2) the data shuffling seed was varied, with fixed weight

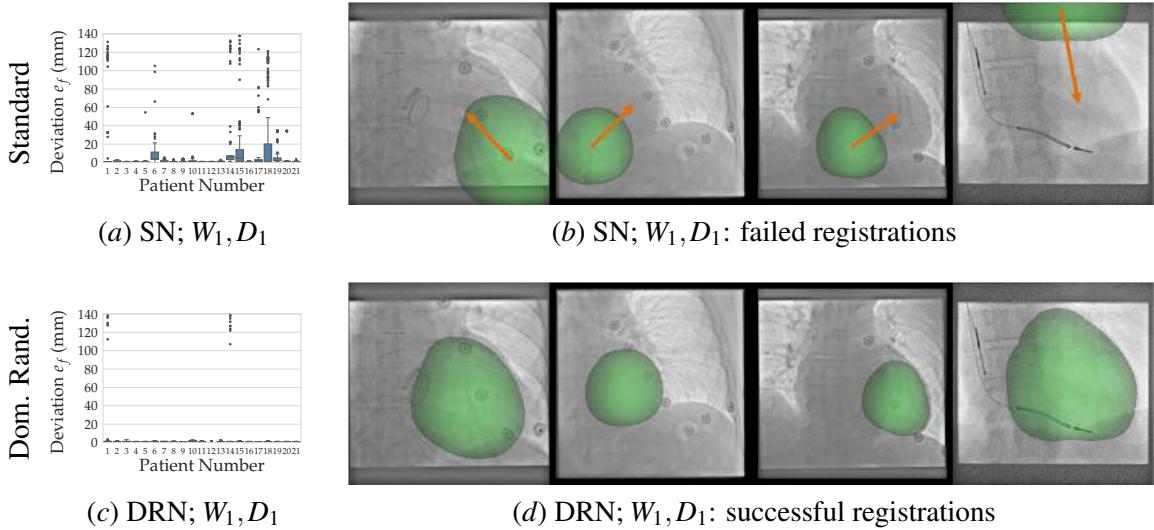


Figure 3: Effects of domain randomization on the target domain (clinical patient data). Baseline quantitative and qualitative results with fixed seeds ( $W_1$  and  $D_1$ ) of the (a-b) standard network (SN) and the (c-d) domain randomized network (DRN).

initialization ( $W_1, D_i$ ). Both experiments had similar outcomes: Results on the synthetic data show that, as expected, all training parameters with both networks (SN and DRN) result in highly similar performance, because no domain transfer is required, see Table 1. In the case of the clinical X-ray data, the SN produces acceptable, but highly variable results. The DRN produces more consistent results, the distributions are compressed and the number of outliers is greatly reduced, see Figure 4 for a selection of setups and Figures 9 and 10 for all results.

If looking at individual cases, such as patients 3 and 8, good robustness and no major variation was present with the SN. The features of these images must be shared between the training and target domains, thus they can be registered without an explicit transfer. Cases such as 12 or 15 vary greatly with varying training parameters. The trained SN networks are in different minima of the optimization space, and the features required for performing well on these cases are learned only in some of the setups. A quantitative summary of the translational and rotational robustness results is displayed in Appendix B.

### 3.4. Ablation Study

To investigate the effects of each DR, they were applied individually in a specific setup ( $W_1, D_4$ ), see Figure 5 and Table 3. The translational robustness distributions have improved most significantly through the DR of the intensity. The DR of the collimation mainly suppresses extreme outliers. Combined DR is the most effective.

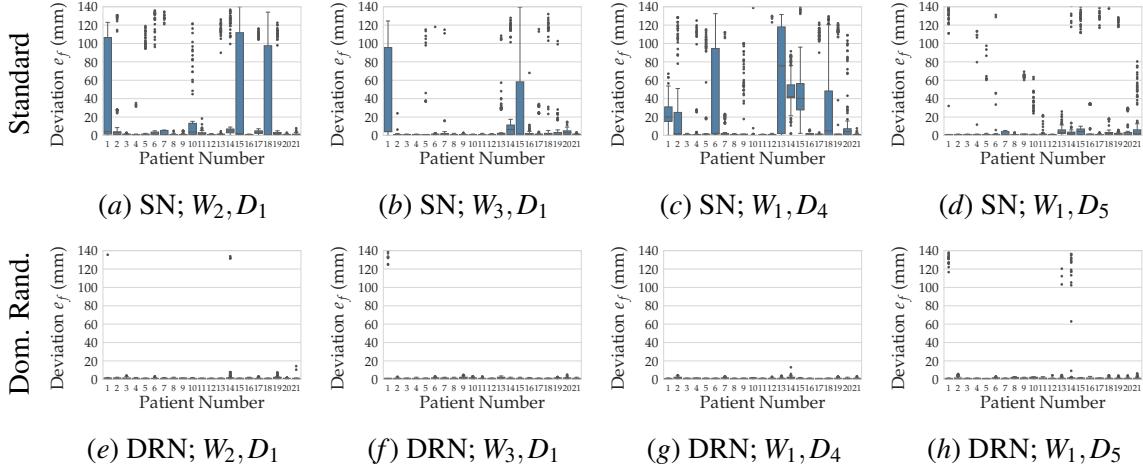


Figure 4: Translational robustness results  $e_f$  of weight initialization ( $W_i$ ) and training data order ( $D_i$ ) variation on clinical patient data. (a–d) Standard network (SN). (e–h) Domain randomized network (DRN).

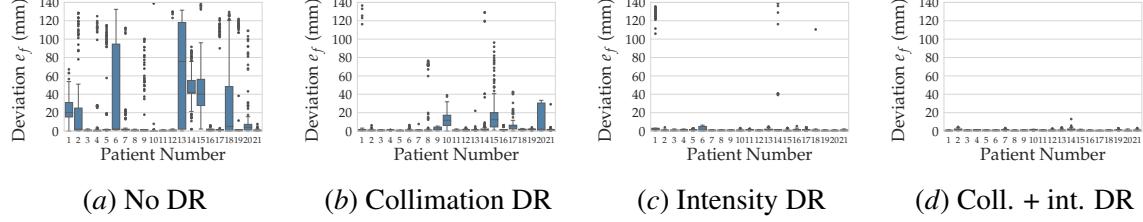


Figure 5: Effects of individual DRs on translational robustness  $e_f$  on clinical data for a single setup ( $W_1, D_4$ ).

#### 4. Conclusion

This paper addresses the performance reduction on clinical data of a cardiac image registration system, if trained only on artificially generated data. Through DR, DRRs are being generated by: 1) varying the intensity transfer function and 2) adding artificial image borders (representing collimation). It was shown, that the trained registration system has higher robustness against geometric perturbations in the test data. Furthermore, it was demonstrated that training with different parameters, i.e., weight initialization or data order, can result in varying results on the target (real) domain, if trained purely on synthetic data. If trained with DR, however, the results are more consistent, there is substantially less variation.

## Acknowledgments

Concepts and information presented are based on research and are not commercially available. Due to regulatory reasons, the future availability cannot be guaranteed. This work was supported by the Wellcome EPSRC Centre for Medical Engineering at KCL (WT 203148/Z/16/Z) and the NIHR Biomedical Research Centre based at GSTT and KCL. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Dept. of Health.

## References

- Tobias Heimann, Peter Mountney, Matthias John, and Razvan Ionasec. Real-time ultrasound transducer localization in fluoroscopy images by transfer learning from synthetic training data. *Medical Image Analysis*, 18(8):1320–1328, 2014. ISSN 13618423. doi: 10.1016/j.media.2014.04.007.
- Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42 (1995):60–88, 2017. ISSN 13618423. doi: 10.1016/j.media.2017.07.005.
- Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. *arXiv*, page 1, 2018. ISSN 0364152X. doi: 10.1107/s00267-013-0043-7.
- Daniel B. Russakoff, Torsten Rohlfing, Kensaku Mori, Daniel Rueckert, Anthony Ho, John R. Adler, and Calvin R. Maurer. Fast Generation of Digitally Reconstructed Radiographs Using Attenuation Fields With Application to 2D-3D Image Registration. *IEEE Transactions on Medical Imaging*, 24(11):1441–1454, 2005. doi: 10.1109/TMI.2005.856749.
- Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. In *IEEE International Conference on Intelligent Robots and Systems*, pages 23–30, 2017. ISBN 1367-8868\n1469-8374. doi: 10.1109/RSS.2017.XIII.034. URL <http://arxiv.org/abs/1611.04201>.
- Daniel Toth, Shun Miao, Tanja Kurzendorfer, Christopher A. Rinaldi, Rui Liao, Tommaso Mansi, Kawal Rhode, and Peter Mountney. 3D/2D model-to-image registration by imitation learning for cardiac procedures. *International Journal of Computer Assisted Radiology and Surgery*, 13(8): 1141–1149, 2018. ISSN 18616429. doi: 10.1007/s11548-018-1774-y. URL <https://doi.org/10.1007/s11548-018-1774-y>.
- Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization. In *arXiv*, pages 1–9, 2018. doi: 10.1109/CVPRW.2018.00143. URL <http://arxiv.org/abs/1804.06516>.
- Mathias Unberath, Jan Nico Zaech, Sing Chun Lee, Bastian Bier, Javad Fotouhi, Mehran Armand, and Nassir Navab. DeepDRR – A Catalyst for Machine Learning in Fluoroscopy-Guided Procedures. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial*

*Intelligence and Lecture Notes in Bioinformatics)*, volume 11073 LNCS, pages 98–106, 2018.  
ISBN 9783030009366. doi: 10.1007/978-3-030-00937-3\_12.

Yue Zhang, Shun Miao, Tommaso Mansi, and Rui Liao. Task driven generative modeling for unsupervised domain adaptation: Application to X-ray image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11071 LNCS, pages 599–607, 2018. ISBN 9783030009335. doi: 10.1007/978-3-030-00934-2\_67.

## Appendix A. Training Curves

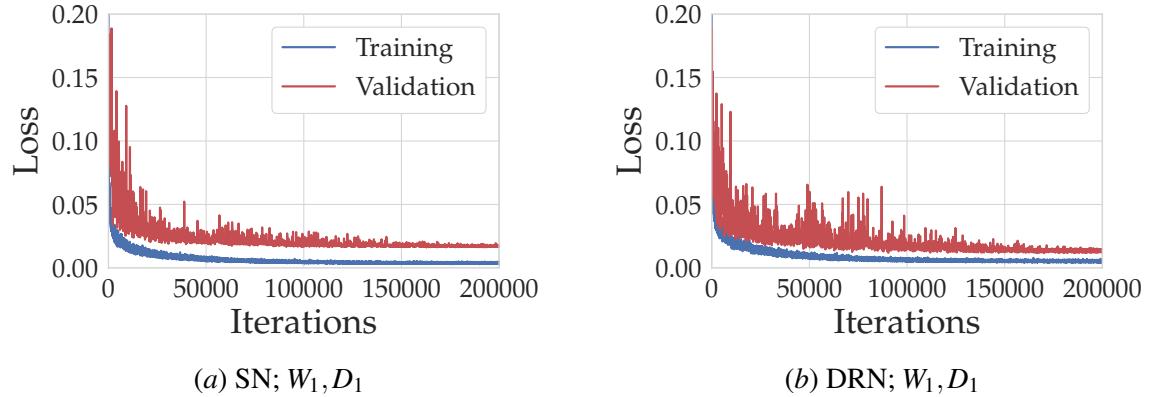


Figure 6: Training and validation curves of the standard (SN) and the domain randomized networks (DRN), with weight initialization and data order seeds,  $W_1$  and  $D_1$  respectively.

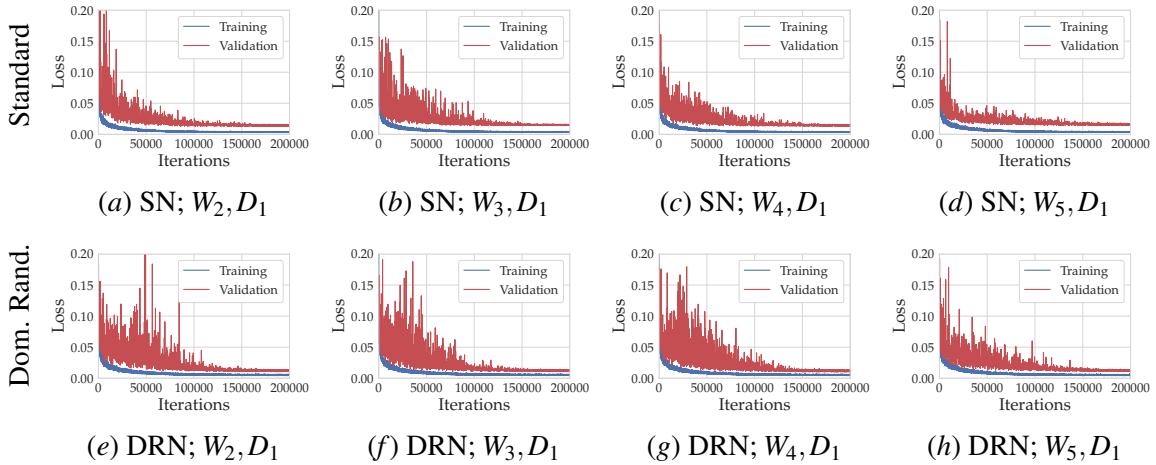


Figure 7: Training and validation curves of the standard (SN) and the domain randomized networks (DRN), with weight initialization varied  $W_i$  and a fixed data order seed  $D_1$ .

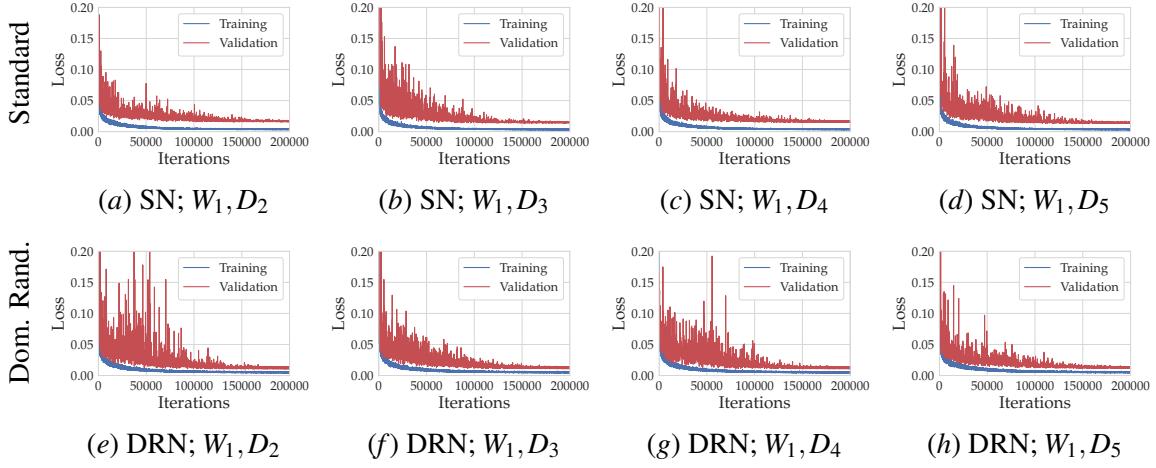


Figure 8: Training and validation curves of the standard (SN) and the domain randomized networks (DRN), with data order varied  $D_i$  and a fixed weight initialization seed  $W_1$ .

## Appendix B. Quantitative Results on Clinical Data

Table 2: Robustness on clinical X-ray data of the standard (SN) and domain randomized networks (DRN) for varying weight initialization  $W_i$  and data order  $D_i$  for translation  $e_f$  and rotation  $\varepsilon_f$ .

	SN/DRN								
	$W_1, D_1$	$W_2, D_1$	$W_3, D_1$	$W_4, D_1$	$W_5, D_1$	$W_1, D_2$	$W_1, D_3$	$W_1, D_4$	$W_1, D_5$
<b>Median <math>e_f</math> (mm)</b>	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
<b>Mean <math>e_f</math> (mm)</b>	5.77/1.38	14.07/1.01	7.30/1.15	10.78/1.34	5.05/1.71	9.42/1.43	6.65/1.41	18.45/0.75	7.60/2.26
<b>75 % <math>e_f</math> (mm)</b>	2.00/1.00	3.00/1.00	1.41/1.00	2.24/1.00	1.41/1.41	1.42/1.00	1.41/1.01	16.55/1.00	2.00/1.41
<b>90 % <math>e_f</math> (mm)</b>	6.32/1.41	97.64/1.41	8.06/1.42	27.00/1.41	5.83/2.24	19.72/1.41	3.16/2.00	85.79/1.41	7.00/2.00
<b>Outliers <math>e_f</math></b>	237/80	263/60	172/28	263/59	154/53	274/54	98/69	342/30	338/114
<b>Median <math>\varepsilon_f</math> (°)</b>	0.46/0.39	0.55/0.50	0.67/0.53	0.60/0.46	0.50/0.49	0.51/0.49	0.52/0.46	0.71/0.46	0.54/0.49
<b>Mean <math>\varepsilon_f</math> (°)</b>	0.99/0.56	3.48/0.82	2.11/0.93	2.35/1.15	1.09/0.71	1.68/0.79	2.01/1.68	2.26/1.06	2.67/1.23
<b>90 % <math>\varepsilon_f</math> (°)</b>	1.69/1.19	7.88/1.55	3.48/1.70	3.57/2.08	1.97/1.47	3.38/1.61	3.37/2.29	5.69/1.48	3.39/1.70

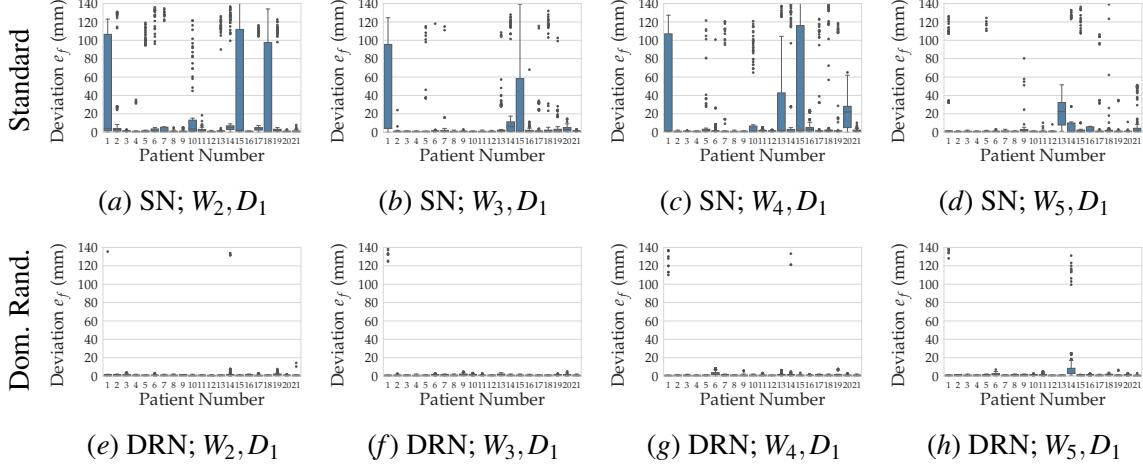


Figure 9: Translational robustness results  $e_f$  on clinical patient data for weight initialization variation ( $W_i$ ) with a fixed training data order seed ( $D_1$ ). (a–d) Standard network (SN). (e–h) Domain randomized network (DRN).

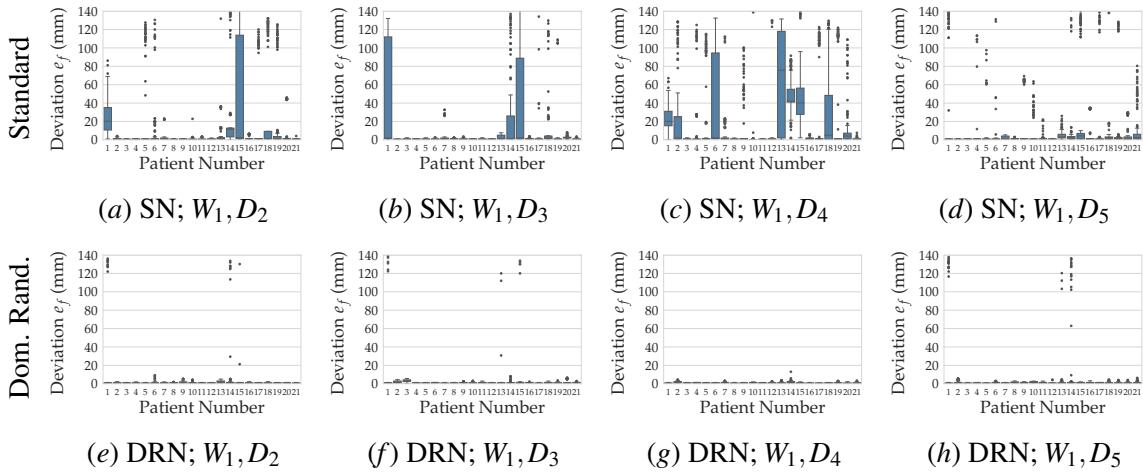


Figure 10: Translational robustness results  $e_f$  on clinical patient data for training data order variation ( $D_i$ ) with a fixed weight initialization seed ( $W_1$ ). (a–d) Standard network (SN). (e–h) Domain randomized network (DRN).

Table 3: Effects of individual domain randomizations on translational  $e_f$  and rotational  $\varepsilon_f$  robustness for a single setup ( $W_1, D_4$ ) on clinical data.

	Translation $e_f$							Angle $\varepsilon_f$ (°)		
	Median (mm)	Mean (mm)	75 % (mm)	90 % (mm)	Outliers	< 5 mm (%)	< 3 mm (%)	Median	Mean	90 %
<b>No DR</b>	1.00	18.45	16.55	85.79	342	69.57	66.78	0.71	2.26	5.69
<b>Collimation DR</b>	1.00	4.03	2.23	7.81	169	84.76	81.09	0.48	0.95	1.78
<b>Intensity DR</b>	1.00	2.42	1.00	1.42	77	97.49	95.10	0.52	0.75	1.54
<b>Coll. + Int. DR</b>	1.00	0.75	1.00	1.41	30	99.94	98.96	0.46	1.06	1.48

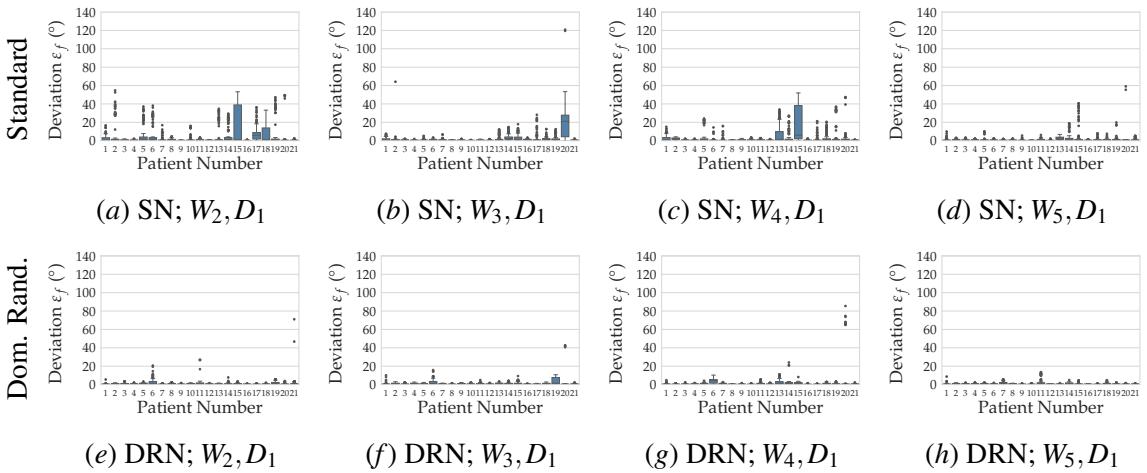


Figure 11: Rotational robustness results  $\varepsilon_f$  on clinical patient data for weight initialization variation ( $W_i$ ) with a fixed training data order seed ( $D_1$ ). (a–d) Standard network (SN). (e–h) Domain randomized network (DRN).

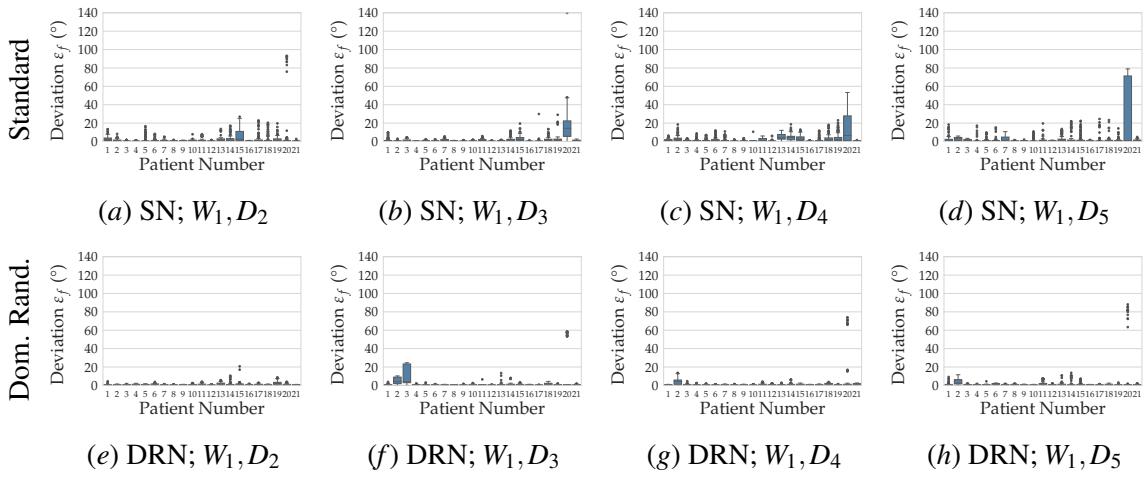


Figure 12: Rotational robustness results  $\epsilon_f$  on clinical patient data for training data order variation ( $D_i$ ) with a fixed weight initialization seed ( $W_1$ ). (a–d) Standard network (SN). (e–h) Domain randomized network (DRN).

# Prediction of Disease Progression in Multiple Sclerosis Patients using Deep Learning Analysis of MRI Data

**Adrian Tousignant<sup>1,3</sup>**

ADRIAN.TOUSIGNANTDURAN@MAIL.MCGILL.CA

**Paul Lemaître<sup>3</sup>**

PLEMAITR@CIM.MCGILL.CA

**Doina Precup<sup>1</sup>**

DPRECUP@CS.MCGILL.CA

**Douglas L. Arnold<sup>2</sup>**

DOUGLAS.ARNOLD@MCGILL.CA

**Tal Arbel<sup>3</sup>**

ARBEL@CIM.MCGILL.CA

<sup>1</sup> School of Computer Science, McGill University, Montreal, Canada

<sup>2</sup> Montreal Neurological Institute, McGill University, Montreal, Canada

<sup>3</sup> Center for Intelligent Machines, McGill University, Montreal, Canada

## Abstract

We present the first automatic end-to-end deep learning framework for the prediction of future patient disability progression (one year from baseline) based on multi-modal brain Magnetic Resonance Images (MRI) of patients with Multiple Sclerosis (MS). The model uses parallel convolutional pathways, an idea introduced by the popular Inception net (Szegedy et al., 2015) and is trained and tested on two large proprietary, multi-scanner, multi-center, clinical trial datasets of patients with Relapsing-Remitting Multiple Sclerosis (RRMS). Experiments on 465 patients on the placebo arms of the trials indicate that the model can accurately predict future disease progression, measured by a sustained increase in the extended disability status scale (EDSS) score over time. Using only the multi-modal MRI provided at baseline, the model achieves an AUC of  $0.66 \pm 0.055$ . However, when supplemental lesion label masks are provided as inputs as well, the AUC increases to  $0.701 \pm 0.027$ . Furthermore, we demonstrate that uncertainty estimates based on Monte Carlo dropout sample variance correlate with errors made by the model. Clinicians provided with the predictions computed by the model can therefore use the associated uncertainty estimates to assess which scans require further examination.

**Keywords:** Deep learning, Multiple Sclerosis, MRI, multi-modal, disease progression

## 1. Introduction

Neurological diseases, such as Alzheimer disease (AD) and Multiple Sclerosis (MS), present major burdens on patients, their families, and on society as a whole (Naci et al., 2010; Brookmeyer et al., 2007). As an example of the financial burden alone, the mean lifetime cost per MS patient has been estimated to be \$2.5 million (Patwardhan et al., 2005). Correctly identifying patients whose condition is at risk of worsening by predicting their future disease course would be enormously beneficial to patients as well as the health care system. More aggressive treatments could be offered to patients for whom the prognosis is poor, setting the stage for personalized medicine. Furthermore, clinical trials would be faster and cheaper if one could accurately predict future disease progression in the early phases of the trial. In this paper, we focus on the particular context of MS, an inflammatory, demyelinating, and degenerative disease of the central nervous system resulting in a range of physical and cognitive disabilities (MS Society of Canada, last accessed on 06/30/18). To date, treatments

have been successfully developed to suppress the acute inflammatory demyelinating lesions that are the hallmark of this disease, and the associated relapses, but these treatments have little or no effect on the relentless disability progression that develops in some patients.

MRIs are an integral part of the clinical management of MS due to their exquisite sensitivity to the development of new MS lesions. MRIs allow clinicians to monitor the evolution of lesions and the effectiveness of drugs being used to suppress them. New lesion formation has also been extremely valuable as a biomarker of treatment efficacy in early stage clinical trials (Barkhof et al., 1997; Frisoni et al., 2010). There have been several successful imaging studies which used MRI biomarkers to determine disability level and progression for a variety of diseases (Jack et al., 2004; Losseff et al., 1996; Ulla et al., 2013). In the context of MS, MRIs have been shown to help predict future lesion activity, defined as the presence of new or enlarged *T2-lesions* in future images, using a *Bag-of-Lesions* brain representation, and to identify potential treatment responders based on carefully designed image features (Doyle et al., 2017). There has also been some success in predicting the conversion to Clinically Definite MS (CDMS) using a SVM on radiomic lesion features (Wottschel et al., 2015) and using a CNN on lesion masks (Yoo et al., 2016). However, although handcrafted features have been successfully adapted to these contexts, there are currently no known imaging biomarkers predictive of future MS disability progression, and it has been shown that clinically-derived lesion features alone are not sufficiently predictive (Zhao et al., 2017). Given the need for data-driven models, deep learning provides an attractive approach to building predictors of disease progression based on MRI, particularly given the recent success of this methodology in a wide variety of tasks in computer vision (Krizhevsky et al., 2012; Deng et al., 2014), and in medical image analysis (Menze et al., 2015; Carass et al., 2017). The capacity of deep learning models to extract predictive features from data is particularly desirable, given that there are currently no clear biomarkers for MS progression.

In this work, we present the *first* automatic, end-to-end deep learning framework for the prediction of future patient disability progression based on multi-modal brain MRI of patients with MS. We use a deep 3D CNN with parallel convolutional pathways, an architecture inspired by the popular Inception net (Szegedy et al., 2015). The model is trained on two large proprietary, multi-scanner, multi-center, clinical trial datasets of patients with Relapsing-Remitting Multiple Sclerosis (RRMS). The model successfully predicts if a significant increase in future clinical disability will occur within a year, with an AUC of  $0.66 \pm 0.055$ . We show that supplementing the MRI with *T2* and *Gadolinium-enhanced* lesion labels (provided by experts) at baseline, if such labels are available, further increases prediction accuracy compared to using MRI alone, resulting in an AUC of  $0.701 \pm 0.027$ . Both proposed models outperform a VGG-like 3D CNN (Simonyan and Zisserman, 2014).

A key ingredient to the successful integration of machine learning methods into clinical workflows is the clinician's trust in the results provided by the learning system. Garnering this trust involves two important steps: (a) providing a quantitative measurement of confidence in the prediction results, and (b) establishing that when the system is confident in its predictions, these are indeed correct. The confidence measurements can then be communicated to the clinicians along with the predictions, so they can determine when further human review or particular attention is required. Hence, our proposed model also provides uncertainty estimates in addition to the predicted progression. This is accomplished by training the deep learning network with dropout, and taking repeated Monte Carlo (MC) samples of the prediction using dropout at test time (Leibig et al., 2016; Nair et al., 2018; Gal and Ghahramani, 2016; Tanno et al., 2017). The sample variance obtained through

this procedure provides a measure of uncertainty in the output. Our results show that, by applying a threshold on the uncertainty level of the output and only considering the patients for which the model is most confident, the prediction performance increases. This suggests a strong correlation between high model uncertainty and incorrect predictions.

## 2. Methodology

We propose a deep learning framework for the prediction of disability progression of MS patients based on MRI, within one year from the baseline scan. The model uses multi-modal MRI volumes as input. The volumes for a patient at baseline are concatenated together and the different modalities are used as input channels. The model consists of three consecutive 3D convolutional blocks, followed by 2 fully connected layers. A dropout layer is used after each convolutional block and fully connected layer. The purpose of these dropout layers is twofold. First, they serve to limit overfitting, which is crucial when working with a small dataset. Secondly, using dropout at test time allows us to take Monte Carlo (MC) samples of predictions, which can be used to quantify the uncertainty of the network in its output. Rectifier linear units are used at all layers except the output, which is a sigmoid.

The standard cross-entropy loss between the progression label in the training data and the model’s prediction is then used to drive the learning process. Note that progression labels are binary and assigned after one year of follow-up visits, in accordance with clinical practice (as described in detail in Sec.3.1). Figure 1 shows a high-level overview of the model.

### 2.1. Convolutional Block

When working with MRI of MS patients, an important problem is the significant discrepancy in the location, shape and size of important lesions, which is in fact part of the difficulty of finding biomarkers based on brain MRI. An important goal in our work was to endow our model with the capability to incorporate information from features of different dimensions in an independent manner. To this end, the deep network is modularized into blocks with parallel convolutional pathways of different focal resolution sizes. Each convolutional block consist of 4 parallel pathways. The first pathway is a max-pooling layer with stride 2, whose purpose is to help propagate of information in a manner similar to residual connections. The 3 other pathways have resolution windows of 3, 5, and 7 respectively. Instead of naively using the expensive  $5 \times 5 \times 5$  and  $7 \times 7 \times 7$  convolution layer, we stack consecutive  $3 \times 3 \times 3$  convolution layers, saving both in memory and in computation. Each convolutional layer has 64 kernels and a stride of 1, except for the last layer of each pathway, which has a stride of 2. The pathways are then concatenated and fed as input to the next block.

## 3. Experiments and Results

We evaluate the ability of the model learned by our approach to predict clinical progression in MS patients through a series of experiments. We first explore the ability of the network to predict future disease progression based on image features extracted from MRI at baseline. To this end, we train our model using 5 different MRI channels per patient, *T1c*, *T1p*, *T2w*, *Pdw*, and *FLAIR*.

The second experiment combines the 5 input MRI modalities with lesion labels masks, including both *T2w* lesion maps and *Gadolinium-enhanced* lesion maps, in order to explore whether this additional information, if available at baseline, improves the model’s accuracy. We compare the

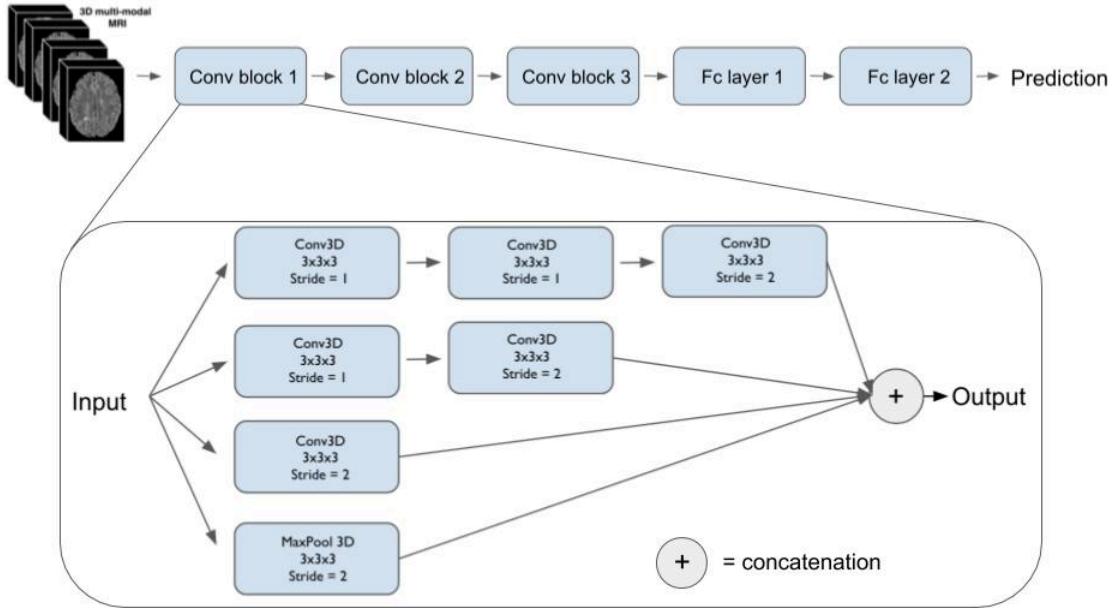


Figure 1: Proposed 3D CNN architecture. MRI volumes are provided as inputs. Information is propagated through three repeated 3D convolution blocks. The last layer is flattened and fed to two fully connected layers. We use ReLUs as the activation function and dropout with a drop probability of 0.5. A zoom-in of the convolutional block unit is provided. Each convolutional block is composed of 4 parallel pathways. The first three consist of different numbers of  $3 \times 3 \times 3$  convolutional layers, while the last is a 3D MaxPooling layer to help with information propagation. There is a stride of 1 for every layer, except for the last layer of each pathway, where the stride is 2.

results of the proposed network to a 3D version of the popular VGG model ([Simonyan and Zisserman, 2014](#)) with a similar number of parameters. Our comparison model uses three series of conv-conv layers with kernel size of  $3 \times 3 \times 3$  and stride of 1 followed by a max pooling layer of size  $2 \times 2 \times 2$  and stride of 2. The number of kernels for each series of conv-conv layers is 64, 256, and 512 respectively. These numbers were chosen to keep the total number of parameters as close as possible to those of our proposed model.

Finally, we explore the effectiveness of using Monte Carlo (MC) sample variance as an estimate of model uncertainty. The objective is to evaluate whether filtering out the most uncertain examples on the ROC plot leads to higher AUC on the ROC curve constructed from the predictions of the remaining examples. If this is indeed the case, then the system is generally correct when certain, so uncertainty measurements are useful to flag problematic cases.

### 3.1. Dataset

The data used for our experiments are drawn from two large proprietary, multi-scanner, multi-center, clinical trial MRI volumes acquired from patients with Relapsing-Remitting MS (RRMS). The image contrast was standardized across sites based on dummy run scans and, once approved, scanners and sequences remained consistent at each site. The MRI volumes were then pre-processed with brain extraction (Smith, 2002), N3 bias field inhomogeneity correction (Sled et al., 1998), Nyúl image intensity normalization (Nyúl and Udupa, 1999), and registration to the MNI-space. Only scans of patients from the placebo arms of both trials were used in the experiments, in order to eliminate the drug effects on our analysis of the natural disease course. In addition, only scans of patients who completed the trial were used. This amounted to scans from a total of 465 patients from two different clinical trials. One trial consisted of 1330 RRMS patients, of which 312 patients were on placebo and completed the trial. Each patient had two MRI scans, taken one year apart, resulting in 624 scans with a non-progression/progression split of 582/42. The second dataset consisted of 543 RRMS patients, of which 153 patients were in the placebo arm and also finished the trial. Patients were scanned at intervals of 24 weeks, resulting in 459 scans with a non-progression/progression split of 398/61. T1-weighted pre- and post- contrast (T1p, T1c), T2-weighted (T2w), Proton Density-weighted (Pdw) and Fluid-attenuated inversion (FLAIR) were used as input modalities for the network. In addition, two lesion masks (T2-weighted and Gadolinium enhanced) were provided with the dataset. The T2-weighted masks were obtained through a semi-manual process in which an automated segmentation algorithm was corrected by a trained expert reader. The Gad-enhancing masks were obtained through consensus between two human experts. In addition to the scheduled scans, patients from both trials had trimestrial clinical follow-up visits, at which the expanded disability status scale (EDSS) score was assessed by a clinician. Each patient was followed by the same clinician over the course of the trial, who was unaware of the treatment assignment in order to prevent bias in the EDSS rating. Binary progression labels were assigned to a scan, in accordance with the clinical definition: change in the expanded disability status scale (EDSS) score was measured over time, and a patient was deemed to progress within a year if the clinical progression criterion was met at least once during the year (see Table 1). A vast majority of patients did not progress during the trial, resulting in a significant class imbalance, with a non-progression to progression ratio close to 9:1. To mitigate the effect of this imbalance, we over-sampled the minority class during training.

Table 1: Requirement for progression given a baseline EDSS score

Definition of Progression	
Baseline EDSS	Criteria
0	An increase of 1.5 or more in EDSS score sustained for 12 weeks or more
0.5 to 5.5	An increase of 1 or more in EDSS score sustained for 12 weeks or more
6.0 and up	An increase of 0.5 or more in EDSS score sustained for 12 weeks or more

### 3.2. Training the network

We trained our network using the RMSProp optimizer with the cross-entropy loss objective (Hinton, 2016). The network was trained for 100 epochs, with a learning rate of  $1e^{-5}$  and a dropout rate of 0.5. Training took approximately 10 hours on a V100 GPU with 16GB of memory. We were forced to use a batch size of 2 only, as we faced memory issues due to the size of the input MRI volumes and the number of model parameters.

### 3.3. Results

The results were evaluated based on the True Positive Rate ( $TPR = \frac{TP}{TP+FN}$ ) and False Positive Rate ( $FPR = \frac{FP}{FP+TN}$ ) receiver operating characteristic (ROC) curve. We obtained the TPR and FPR points by varying the threshold used to binarize the output of the model. These metrics were chosen to assess the performance instead of accuracy due to the large class imbalance. The dataset was split into training (75%), validation (15%) and test sets (10%).  $K$ -fold cross-validation was performed with  $K = 4$ ; the choice of  $K$  was determined based on the size of the dataset. We ensured that data from all the follow-up visits for the same patient was always kept in the same fold, in order to not contaminate the validation and testing sets with training data. As recall and precision are both important for progression prediction, the F-score was used for early stopping.

When only the five MRI modalities were provided as input, the network attained an average area under the curve (AUC) of  $0.66 \pm 0.055$ . With the addition of T2 lesion and Gad-enhancing lesion masks as inputs, the progression prediction AUC improved substantially and the variability across folds was reduced. The AUC for this case was  $0.701 \pm 0.027$ . Using McNemar's test for comparison of classification learning algorithms, we can reject the null hypothesis that our model is similar to a model trained with scrambled labels with a p-value of  $4.84e^{-6}$ . Figure 2 shows the ROC curves for both experiments. Table 2 summarizes the quantitative results for both experiments against the baseline VGG-like 3D CNN which was included for comparison. Both experiments performed significantly better than the baseline, which had an AUC of  $0.615 \pm 0.053$ .

Table 2: Comparison of the model's performance

Networks	# of Parameters	AUC $\pm$ std
VGG-like 3D CNN	13.5M	$0.615 \pm 0.053$
Proposed 3D CNN	14M	$0.66 \pm 0.055$
<b>Proposed 3D CNN with lesion masks</b>	<b>14M</b>	<b><math>0.701 \pm 0.027</math></b>

We quantified the uncertainty of our model's output by taking Monte Carlo dropout prediction samples at test time. We ran 20 forward passes of our model with a dropout rate of 0.5 and calculated the mean and sample variance of the generated values, representing the final prediction output and the associated model uncertainties, respectively. To assess the relationship between our network's confidence in its prediction and its performance, we generated the overall ROC curve (based on the entire set of results) and compared it against the ROC curves showing the results in which the top  $n^{th}\%$  most uncertain predictions were removed. Figure 3 shows that excluding even a few of the most uncertain predictions can improve the model's overall AUC on the remaining predictions. Specifically, the reference curve has an AUC of 0.649. Removing the 2% most uncertain examples is enough to result in an increase in AUC to 0.659. The trend continues when removing the top

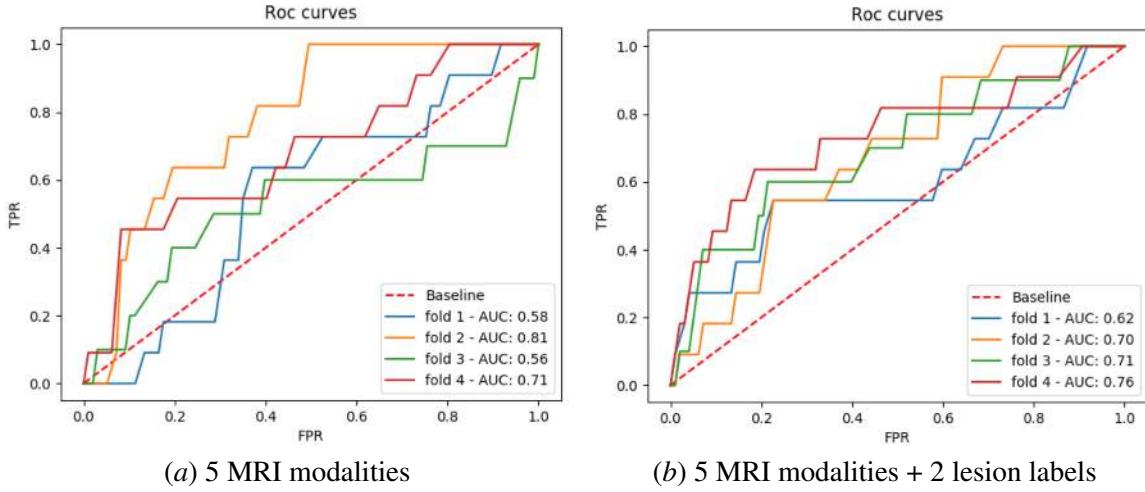


Figure 2: TPR vs FPR ROC curves: (a) Using the 5 MRI modalities only as inputs. AUC of  $0.66 \pm 0.055$  (b) Supplementing the model with the 2 provided lesion label masks. AUC of  $0.701 \pm 0.027$ . Results for each fold of the cross-validation are shown.

10% most uncertain results, which leads to an AUC of 0.681 on the remaining predictions, an improvement of almost 5%.

#### 4. Conclusion and Future Work

In this paper, we develop an end-to-end 3D CNN with parallel convolutional layers for predicting future progression in MS patients using MRI images and clinical assessment of disability. Our results also indicate that supplementing the model with *T2* and *Gad-enhancing* lesion labels, should they be available, further improves prediction accuracy. Finally, our model includes an uncertainty estimate based on MC dropout sample variance. Filtering out very uncertain examples leads to improved results for the remaining predictions. While this work focused on MC sample variance as an uncertainty metric, exploring alternative ways of quantifying uncertainty would be of interest. Additionally, adapting our architecture to leverage longitudinal clinical information (e.g. age, disability stage) could help improve predictions. Finally, exploring ways to interpret the predictions of the model and identify which regions of the brain contributed to the final decisions could help uncover new MS biomarkers, guiding the way of future research and furthering our understanding of the disease.

#### Acknowledgments

This work was supported by an award from the International Progressive MS Alliance (PA-1603-08175).

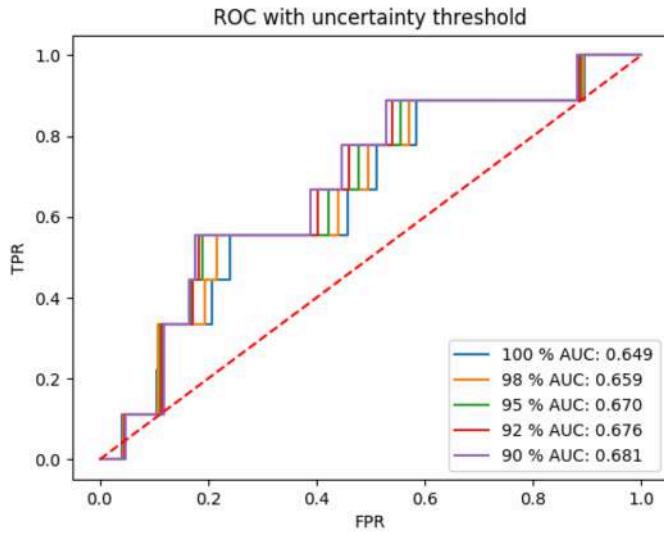


Figure 3: TPR vs. FPR of retained test time predictions when thresholding based on the MC sample variance. The percentage of predictions retained at each curve’s uncertainty threshold is indicated in the color coded legend as well as the corresponding AUC. The reference curve (100%) performance, when no uncertainty thresholding is performed, is also shown (blue).

## References

- Frederik Barkhof, Massimo Filippi, David H. Miller, Philip Scheltens, Adriana Campi, Chris H. Polman, Giancarlo Comi, Herman J. Adèr, Nick Losseff, and Jacob Valk. Comparison of mri criteria at first presentation to predict conversion to clinically definite multiple sclerosis. *Brain*, 120(11):2059–2069, 11 1997.
- Ron Brookmeyer, Elizabeth Johnson, Kathryn Ziegler-Graham, and H Michael Arrighi. Forecasting the global burden of alzheimer’s disease. *Alzheimer’s & dementia*, 3(3):186–191, 2007.
- Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.
- Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.
- Andrew Doyle, Doina Precup, Douglas L Arnold, and Tal Arbel. Predicting future disease activity and treatment responders for multiple sclerosis patients using a bag-of-lesions brain representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–194. Springer, 2017.

- Giovanni B. Frisoni, Nick C. Fox, Clifford R. Jack, Philip Scheltens, and Paul M. Thompson. The clinical use of structural mri in alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML’16, pages 1050–1059. JMLR.org, 2016.
- Geoffrey Hinton. Neural networks for machine learning. Coursera, 2016.
- C. R. Jack, M. M. Shiung, J. L. Gunter, P. C. O’Brien, S. D. Weigand, D. S. Knopman, B. F. Boeve, R. J. Ivnik, G. E. Smith, R. H. Cha, E. G. Tangalos, and R. C. Petersen. Comparison of different mri brain atrophy rate measures with clinical disease progression in ad. *Neurology*, 62(4):591–600, 2004.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- Christian Leibig, Vaneeda Allken, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *bioRxiv*, 2016.
- N. A. Losseff, S. L. Webb, J. I. O’Riordan, R. Page, L. Wang, G. J. Barker, P. S. Tofts, W. I. McDonald, D. H. Miller, and A. J. Thompson. Spinal cord atrophy and disability in multiple sclerosis: A new reproducible and sensitive MRI method with potential to monitor disease progression. *Brain*, 119(3):701–708, 1996.
- Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2015.
- MS Society of Canada. What is MS?, last accessed on 06/30/18. <https://mssociety.ca/about-ms/what-is-ms>.
- Huseyin Naci, Rachael Fleurence, Julie Birt, and Amy Duhig. Economic burden of multiple sclerosis. *Pharmacoeconomics*, 28(5):363–379, 2010.
- Tanya Nair, Doina Precup, Douglas L. Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 655–663, Cham, 2018. Springer International Publishing.
- László G Nyúl and Jayaram K Udupa. On standardizing the mr image intensity scale. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(6):1072–1081, 1999.
- MB Patwardhan, DB Matchar, GP Samsa, DC McCrory, RG Williams, and TT Li. Cost of multiple sclerosis by level of disability: a review of literature. *Multiple Sclerosis Journal*, 11(2):232–239, 2005.

- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.
- Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- Ryutaro Tanno, Daniel E Worrall, Aurobrata Ghosh, Enrico Kaden, Stamatios N Sotiroopoulos, Antonio Criminisi, and Daniel C Alexander. Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–619. Springer, 2017.
- Miguel Ulla, Jean Marie Bonny, Lemlih Ouchchane, Isabelle Rieu, Beatrice Claise, and Franck Durif. Is r<sub>2</sub>\* a new mri biomarker for the progression of parkinson’s disease? a longitudinal follow-up. *PLoS One*, 8(3):e57904, 2013.
- V Wottschel, DC Alexander, PP Kwok, DT Chard, ML Stromillo, N De Stefano, AJ Thompson, DH Miller, and O Ciccarelli. Predicting outcome in clinically isolated syndrome using machine learning. *NeuroImage: Clinical*, 7:281–287, 2015.
- Youngjin Yoo, Lisa W Tang, Tom Brosch, David KB Li, Luanne Metz, Anthony Traboulsee, and Roger Tam. Deep learning of brain lesion patterns for predicting future disease activity in patients with early symptoms of multiple sclerosis. In *Deep Learning and Data Labeling for Medical Applications*, pages 86–94. Springer, 2016.
- Yijun Zhao, Brian C Healy, Dalia Rotstein, Charles RG Guttmann, Rohit Bakshi, Howard L Weiner, Carla E Brodley, and Tanuja Chitnis. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS One*, 12(4):e0174866, 2017.

# Learning beamforming in ultrasound imaging

**Sanketh Vedula<sup>1</sup>\***

**Ortal Senouf<sup>1</sup>\***

**Grigoriy Zurakhov<sup>1</sup>**

**Alex Bronstein<sup>1</sup>**

**Oleg Michailovich<sup>2</sup>**

**Michael Zibulevsky<sup>1</sup>**

<sup>1</sup> *Technion, Israel*

<sup>2</sup> *University of Waterloo, Canada*

SANKETH@CS.TECHNION.AC.IL

SENOUF@CAMPUS.TECHNION.AC.IL

GRISHAZ@CAMPUS.TECHNION.AC.IL

BRON@CS.TECHNION.AC.IL

OLEGM@UWATERLOO.CA

MZIB@CS.TECHNION.AC.IL

## Abstract

Medical ultrasound (US) is a widespread imaging modality owing its popularity to cost efficiency, portability, speed, and lack of harmful ionizing radiation. In this paper, we demonstrate that replacing the traditional ultrasound processing pipeline with a data-driven, learnable counterpart leads to significant improvement in image quality. Moreover, we demonstrate that greater improvement can be achieved through a learning-based design of the transmitted beam patterns simultaneously with learning an image reconstruction pipeline. We evaluate our method on an in-vivo first-harmonic cardiac ultrasound dataset acquired from volunteers and demonstrate the significance of the learned pipeline and transmit beam patterns on the image quality when compared to standard transmit and receive beamformers used in high frame-rate US imaging. We believe that the presented methodology provides a fundamentally different perspective on the classical problem of ultrasound beam pattern design.

**Keywords:** Ultrasound Imaging, Deep Learning, Beamforming

## 1. Introduction

Recently, there has been a surge of interest in applying learning-based techniques to improve ultrasound imaging. In (Senouf et al., 2018) and (Vedula et al., 2018), we demonstrated that convolutional neural networks (CNNs) can be employed to reconstruct high-quality images acquired through high-framerate ultrasound acquisition protocols. Similarly, in (Gasse et al., 2017), the authors proposed that CNNs could be used as a means to perform plane-wave compounding requiring significantly lesser number of plane-waves to reconstruct a high-quality image. (Simson et al., 2018) proposed to approximate time-consuming beamformers such as minimum-variance beamforming using CNNs. In (Luchies and Byram, 2018), the authors proposed to use process time-delayed and phase-rotated signals using fully connected networks showing to improve ultrasound image reconstruction. Apart from ultrasound image formation, CNNs were used in ultrasound post-processing for real-time despeckling and CT-quality image reconstruction (Vedula et al., 2017), for speed-of-sound estimation (Feigin et al., 2018) and for ultrasound segmentation directly from the raw-data (Nair et al., 2018).

---

\* Contributed equally

**Contributions.** Viewing US imaging as an inverse problem, in which a latent image is reconstructed from a set of measurements, the above mentioned studies focused on learning (parts of) the inverse operator producing an image from the measurements. The scope of the present paper differs sharply in the sense that we propose to learn the parameters of the *forward model*, specifically, the transmitted patterns. We propose to jointly learn the end-to-end transmit (Tx) and receive (Rx) beamformers optimized for the task of high-framerate ultrasound imaging, in which the number of measurements per image has a direct impact on the frame rate. We demonstrate a significant improvement in the image quality compared to the standard patterns used in this setting.

Unlike our previous works ([Senouf et al., 2018](#); [Vedula et al., 2018](#)) that train separate networks for the *in-phase* (I) and *quadrature* (Q) components of the demodulated received ultrasound data, we propose a unified dual-pathway network that trains jointly I and Q minimizing for the loss defined on the final envelope image (Figure 1). We also propose a new beamforming layer inspired by ([Jaderberg et al., 2015](#)), that implements beamforming as a *differentiable* geometric transformation between pre-beamformed Rx signal and the beamformed one. This results in a fully-differentiable end-to-end Rx beamforming and signal processing pipeline that can be easily generalized to a variety of imaging settings. By rendering the end-to-end Rx pipeline differentiable, we demonstrate that the Tx protocols can be optimized together with the Rx beamforming and reconstruction pipeline, leading to significant improvement in image quality. To the best of our knowledge, this is the first time simultaneous end-to-end learning of hardware parameters and signal processing algorithms are used in US imaging.

## 2. Methods

Traditionally, a US imaging pipeline consists of the following stages: Tx beamforming, acquisition, Rx beamforming, and image formation. In Tx beamforming, depending on the desired frame-rate and quality, a suitable number of transmissions and their corresponding beam profile are chosen and the piezo-electric transducers are programmed accordingly to transmit the beams. Post-transmission, the echoes are received by the same transducer array; these signals are demodulated and focused by applying the appropriate time-delays and phase-rotations to produce the beamformed signal. The beamformed signal is further processed to correct the artifacts (if acquired through high frame-rate transmit modes) and apodized to suppress the side-lobes. We refer to these stages of processing the demodulated signals collectively as *Rx beamforming* (Figure 1). After Rx beamforming, the envelope is extracted from the complex signal, followed by a log-compression and scan-conversion to produce the final ultrasound image.

### 2.1. Learned end-to-end Rx pipeline

In our previous studies ([Senouf et al., 2018](#); [Vedula et al., 2018](#)), we have used a symmetric encoder-decoder multi-resolution neural network in order to fix the distorted received US signal and get the higher quality undistorted signal. Two networks were trained separately for the I and Q signals, mostly due to computational and technical difficulties to train one network for both. In this paper, we present an architecture that comprises two separate paths for I and Q followed by a layer forming the envelope signal, on which the loss is calculated.  $\Theta_I$  and  $\Theta_Q$  in Figure 2 denote the parameters of the two encoder-decoder networks with an architecture similar to that of a U-Net ([Ronneberger et al., 2015](#)). Moreover, in our previous works we have trained and applied the networks to the time-delayed and phase rotated signals, which would not allow us to perform manipulations on

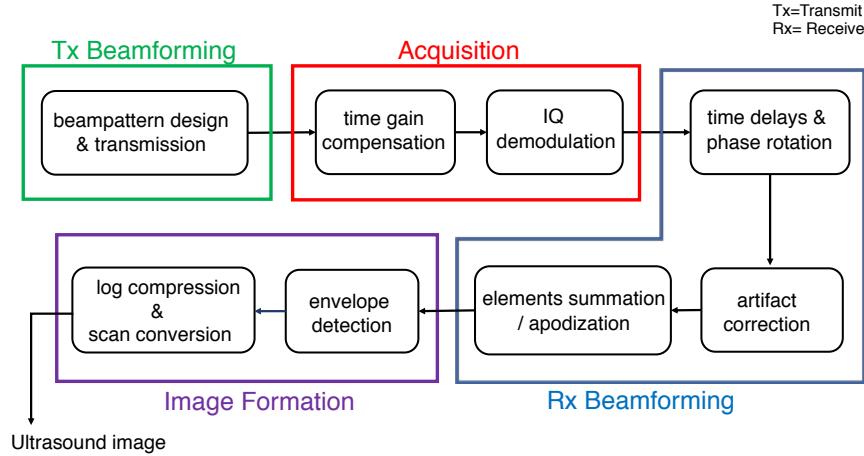


Figure 1: The ultrasound imaging pipeline

transmission (Tx) patterns. In this work, we have implemented a time-delays and phase rotation stage (referred to as *dynamic focusing*) in the network architecture, which allows to work on the pre-Rx-beamformed signals directly, as described in Figure 2.

Performing time-delays and phase-rotations through convolutions is not trivial because it would require a very large support of surrounding data points. This, in turn would require a computationally intractable number of arithmetic operations to approximate the delays. In order to overcome this problem, we propose to perform time-delays and phase-rotations as a differentiable geometric transformation of the pre-beamformed signal. We introduce a spatial transformation layer inspired by the works of (Jaderberg et al., 2015) and (Skafte Detlefsen et al., 2018), in which the authors proposed a differentiable sampling and interpolation method in order to train and apply affine and, more generally, diffeomorphic transformations to the input. Here, we apply the explicit time delays and phase-rotation (*dynamic focusing*) in a similar fashion. Given the raw signal  $\phi_m(t, \alpha)$  corresponding to focused beams direction  $\alpha$  read out from the  $m$ -th array element at location  $\delta_m$  and time  $t$ , we construct the time-delayed signal as  $\hat{\phi}_m(t, \alpha) = \phi_m(\hat{t}, \alpha)$ , where

$$\hat{t} = \frac{t}{2} + \sqrt{\frac{t^2}{4} - t \sin \alpha \frac{\delta_m}{c} + \left( \frac{\delta_m}{c} \right)^2},$$

and  $c$  is the speed of sound in the tissue, assumed to be 1540 m/s. In addition, in order to eliminate phase error, phase rotation is applied to the complex signal in its explicit form, as described in (Chang et al., 1993):

$$\begin{pmatrix} \Re \text{IQ} \\ \Im \text{IQ} \end{pmatrix} = \begin{pmatrix} \cos(\omega_0(\hat{t}-t)) & -\sin(\omega_0(\hat{t}-t)) \\ \sin(\omega_0(\hat{t}-t)) & \cos(\omega_0(\hat{t}-t)) \end{pmatrix} \begin{pmatrix} \Re \hat{\phi}_m(t, \alpha) \\ \Im \hat{\phi}_m(t, \alpha) \end{pmatrix},$$

where  $\omega_0$  is the modulation frequency and  $\Re$  and  $\Im$  denote, respectively, the real and imaginary parts of a complex number.

The *dynamic focusing* is placed after the *Tx beamformer* layer and before the reconstruction network, as depicted in Figure 2. While in our implementation, all the parameters defining the time delay and phase rotation transformations are fixed, they can be trained as well.

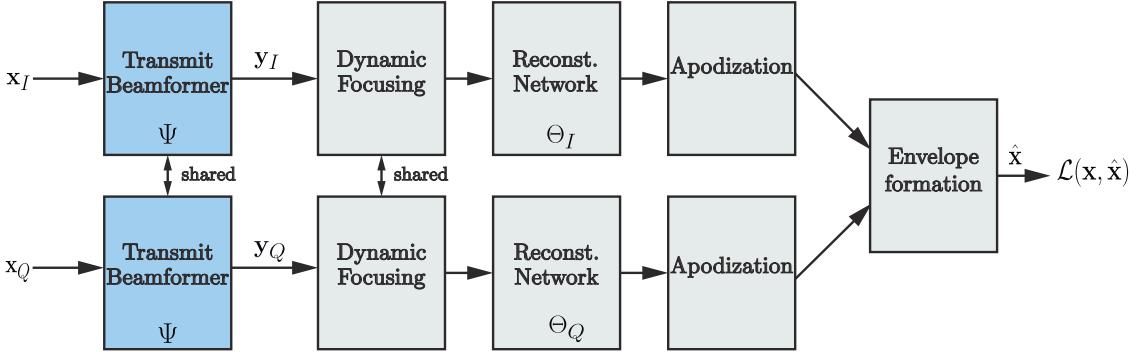


Figure 2: Learned end-to-end Tx-Rx pipeline. The stages: dyanamic focusing, reconstruction network, apodization and envelope formation are together referred to as Rx beamforming.

## 2.2. Learning optimal transmit patterns

The problem of learning optimal transmitted patterns together with Rx beamforming and reconstruction can be formulated as a simultaneous learning of the forward model and its (approximate) inverse. Ultrasound imaging can be viewed end-to-end as a process that given a latent image  $x$  (the object being imaged) generates a set of measurements  $y$  thereof by sampling from a parametric conditional distribution  $y \sim p_\psi(y|x)$ . This conditional distribution is known as the likelihood in the Bayesian jargon, and can be viewed as a stochastic forward model. The set of parameters  $\psi$  denotes collectively the settings of the imaging hardware, including the patterns transmitted to obtain the measurements.

The goal of the signal processing pipeline is to produce the an estimate  $\hat{x}$  of the latent image  $x$  given the measurements  $y$ . We denote the estimator as  $\hat{x}_\theta(y)$  and refer to it as the inverse operator, implying that it should invert the action of the forward model. The set of parameters  $\theta$  denotes the trainable degrees of freedom of the reconstruction pipeline; in our case, these are the weights of the reconstruction neural network. We propose to simultaneously learn the parameters of both the forward model and the inverse operator such as to optimize performance in a specific task. This can be carried out by minimizing the expected loss,

$$\min_{\theta, \psi} \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \mathbb{E}_{\mathbf{y} \sim p_\psi(\mathbf{y}|\mathbf{x})} \mathcal{L}(\hat{\mathbf{x}}_\theta(\mathbf{y}), \mathbf{x}),$$

where  $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x})$  measures the discrepancy between the ground truth image  $\mathbf{x}$  and its estimate  $\hat{\mathbf{x}}$ . In practice, the expectations are replaced by finite-sample approximation on the training set. Note that the expectation taken over  $\mathbf{y} \sim p_\psi(\mathbf{y}|\mathbf{x})$  embodies the parametric forward model whose parameters  $\psi$  (reflecting the transmission pattern) are optimized simultaneously with the parameters of the inverse operator (i.e., the computational process applied to the measurement  $\mathbf{y}$  to recover the latent signal), in our case, the reconstruction network. This training regime resembles in spirit the training of autoencoder networks; in our case, the architecture of the encoder is fixed as dictated by the imaging hardware, and only parameters under the user's control can be trained.

The idea of simultaneously training a signal reconstruction process and some parameters of the signal acquisition forward model has been previously corroborated in computational imaging,

including compressed tomography (Menashe and Bronstein, 2014), phase-coded aperture extended depth-of-field and range image sensing (Haim et al., 2018). In all the mentioned cases, a significant improvement in performance was observed both in simulation and in real systems.

In our current work, we refer only to first harmonic ultrasound imaging, whose forward model is linear. This means that applying manipulations to the received signal is equivalent to applying them on the transmitted signal, as has been shown in (Prieur et al., 2013). This way the forward model is parameterized by a set of linear combinations of the original received beam,

$$\mathbf{y}_j = \sum_{i=1}^L \psi_{ij} \mathbf{x}_i, \quad \{\mathbf{y}_j\}_{j=1}^M$$

where  $L$  is the number of the original received beams,  $M$  is the number of new learned beams, and the matrix  $\psi$  encodes the transmit beam patterns. It has been shown (Prieur et al., 2013) that this approach can faithfully emulate measurements that would be formed from a more complex excitation.

### 3. Experiments and discussion

#### 3.1. Data acquisition

The FOV was scanned by 140/140 Tx/Rx lines, each of them covered a sector of  $0.54^\circ$ . We refer to this baseline acquisition scenario as *single-line acquisition* (SLA) and consider it to be the ground truth in all reduced transmission experiments. In order to assess the generalization performance of our method, we used a cine loop from a patient whose data were excluded from the training/validation set.

#### 3.2. Settings

In order to evaluate the contribution of the joint training of the transmit pattern and the received signal reconstruction, we have designed a two-stage experiment. First, we trained only the reconstruction network and fixed the Tx beamforming parameters. Second, we used a pre-convergence checkpoint of the reconstruction network as a starting point for the joint training. At this stage, we also trained the Tx parameters. In order to factor out the influence of the optimization algorithm, we trained the reconstruction network in both stages with the same optimizer (Adam, initial learning rate = 0.005). The Tx parameters were trained using the momentum optimizer with a decaying learning rate (initial learning rate = 0.005). The loss function,  $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x})$ , was set to the  $L_1$  error.

**Different initializations.** We performed the two-stage experiment with different initializations for the Tx parameters using known reduced transmission methods as well as random initialization. We fixed the decimation factor to 10, meaning that instead of the 140 original acquisitions, only 14 measurements were emulated and provided to the reconstruction network. One initialization method was the *multi-line acquisition* (MLA) in which for every wide transmitted beam, 10 (as the decimation factor) Rx narrow beams are reconstructed. Each 10–MLA acquisition is emulated by averaging over 10 consecutive *single-line acquisition* (SLA) Rx signals (as depicted in Figure 12 in the Appendix) (Rabinovich et al., 2013). Another initialization method is the *multi-line transmission* (MLT) in which a comb of uniformly spaced narrow beams is transmitted simultaneously. Each 10–MLT acquisition is emulated by summing over 10 uniformly spaced received Rx signals from

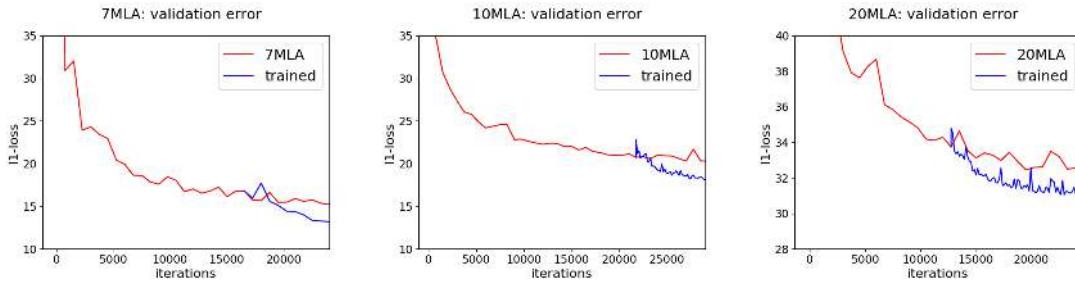


Figure 3: Convergence plots. Depicted from left to right are the validation error plots of 7–, 10– and 20–MLA. The red and blue lines indicate the *learning Rx* and *learning Tx-Rx* settings, respectively.

SLA (as presented in Figure 12, in the Appendix) (Rabinovich et al., 2015). Finally, a random initialization was used to emulate, in a way, a plane wave excitation (Montaldo et al., 2009), in which there is no directivity to the beam pattern. In this experiment, mentioned in this paper as 10–random, 14 acquisitions of distinct random patterns were emulated.

**Different decimation rates.** In this experiment, we fixed the initialization to MLA and performed the above described two-stage experiment over different decimation rates 7, 10, and 20.

### 3.3. Results and discussion

**Notation.** For all the experiments presented within the paper, *Learned Rx* refers to the setting where the transmission is fixed and the reconstruction network alone is trained and *Learned Tx-Rx* refers to the setting in which the transmission patterns are jointly learned with the reconstruction network. *Fixed Tx – DAS* refers to the setting where the fixed transmissions are beamformed using a standard delay-and-sum (DAS) beamformer, and *Learned Tx – DAS* is the setting where learned transmissions are beamformed using a delay-and-sum Rx beamformer.

**Convergence.** Figure 3 displays the validation error plot of the two stages training for the different decimation rates experiment. Each iteration corresponds with a mini-batch, which in our settings its size has been set to one. The error gap between the learned-Rx and the jointly learned Tx-Rx, in favour of the latter, supports our claim for the superiority of joint learning of forward and inverse models in the case of US acquisition. A similar behaviour was observed for other initializations.

**Train + test split.** We generated a dataset for training the network using cardiac data from six patients; each patient contributed 4 – 5 cine loops containing 32 frames each. The networks were trained on the cineloops of five patients and the testset consists of the cineloops from the patient that was excluded from the trainset. The total trainset consisted of 745 frames, while the testset consisted of 160 frames.

**Quantitative results.** We present the quantitative evaluation of the first cineloop (32 frames) in Table 1, the quantitative results for the rest of the cineloops are summarized in the supplementary material<sup>1</sup>. Table 1 (top), summarizing the average quality measures for the different decimation

1. Supplementary material: available [here](#).

	7-MLA			10-MLA			20-MLA		
	PSNR	SSIM	L1-error	PSNR	SSIM	L1-error	PSNR	SSIM	L1-error
Fixed Tx – DAS	33.76	0.955	–	32.34	0.941	–	29.6	0.91	–
Learned Tx – DAS	34.03	0.96	–	32.73	0.95	–	29.87	0.916	–
Learned Rx	42.56	0.987	19.14	39.56	0.975	24.31	35.02	0.924	38.36
Learned Tx-Rx	43.4	0.99	15.94	39.98	0.98	22.19	35.32	0.95	36.24

	10-MLA			10-MLT			10-random		
	PSNR	SSIM	L1-error	PSNR	SSIM	L1-error	PSNR	SSIM	L1-error
Fixed Tx – DAS	32.34	0.941	–	24.39	0.855	–	24.26	0.865	–
Learned Tx – DAS	32.73	0.95	–	25.22	0.878	–	25.34	0.88	–
Learned Rx	39.56	0.975	24.31	33.66	0.92	47.99	34.7	0.935	46.7
Learned Tx-Rx	39.58	0.98	22.19	35.04	0.92	41	36.52	0.95	38

Table 1: Comparison of average PSNR, SSIM and  $L_1$  error measures between different decimation rates of transmissions (top) and different initializations (bottom). First and second rows indicate the performance of fixed and learned transmissions with a standard delay-and-sum (DAS) beamformer, respectively. Third and fourth rows indicate the results corresponding to learned Rx and learned Tx-Rx experiment settings, respectively.

rates, shows improved performance in the sense of the  $L_1$  error used to train the models, and in the sense of the peak signal-to-noise ratio (PSNR), which is correlated to the  $L_1$  loss. It is interesting to observe that an improvement was also observed in the sense of the structure-similarity (SSIM) measure, for which the models were not trained. In Tables 1 and 2, we can observe that the learned Rx pipeline performs significantly better than the fixed Tx with a DAS beamformer. Similar behavior can be observed in all the experiments. More interestingly, one can see that the learned transmissions perform better than the fixed ones even with the DAS beamformer. The best performance, with a significant margin, is achieved when the transmit patterns and the Rx beamformer are jointly learned, in all settings. Comparison between different initializations of transmission patterns for a fixed decimation factor is presented in Table 1 (bottom). Observe that the transmission pattern initialized with MLA performs better than MLT and random initializations, also by a significant margin.

Visual inspection of the results of the two-stage training experiment for both different rates and different initializations settings, on one of the test frames is displayed in Figures 10, 11 in the Appendix, along with the corresponding difference images (compared to SLA) and contrast (Cr), and contrast-to-noise (CNR) ratios (Tables 3, 4). These results suggest a better interpretability of the images generated from the jointly trained Tx-Rx models, especially for higher decimation rates (as displayed for the 20-MLA initialization in Figure 4) and the less-directed initializations (MLT and random).

**Generalization to phantom dataset.** A phantom dataset consisting of 46 frames was acquired with the same acquisition setup as of the cardiac dataset from a tissue mimicking phantom (GAM-MEX Ultrasound 403GS LE Grey Scale Precision Phantom). In order to evaluate the generalization performance of the proposed approach, we test all the networks that were originally trained on the

cardiac samples on the phantom dataset. Results in Table 2 suggest that the proposed methodology while being trained on the cardiac data, generalizes well to the phantom, which is also consistent with the observations we made in our previous works (Vedula et al., 2018; Senouf et al., 2018). Firstly, this indicates that our reconstruction CNN does not overfit to the anatomy it was trained on. Secondly, and more interestingly, we can observe that the *Learned Tx-Rx* setting consistently outperforms the *Learned Rx* setting, which indicates that the transmit patterns learned over the cardiac data also transfer well to the phantoms.

	7-MLA			10-MLA			20-MLA		
	PSNR	SSIM	L1-error	PSNR	SSIM	L1-error	PSNR	SSIM	L1-error
Learned Rx	40.92	0.978	5.3	32.09	0.911	9.91	30.23	0.901	12.94
Learned Tx-Rx	43.73	0.989	3.35	31.14	0.92	7.62	31.92	0.903	10.53

	10-MLA			10-MLT			10-random		
	PSNR	SSIM	L1-error	PSNR	SSIM	L1-error	PSNR	SSIM	L1-error
Learned Rx	32.09	0.911	9.91	31.05	0.624	15.91	31	0.667	16.34
Learned Tx-Rx	31.14	0.92	7.62	32.098	0.711	13.765	31	0.76	14.276

Table 2: Generalization to phantom dataset. Comparison of average PSNR, SSIM and  $L_1$  error measures between different decimation rates of transmissions (top) and different initializations (bottom). Top and bottom rows indicate the results corresponding to learned Rx and learned Tx-Rx experiment settings, respectively.

**Learned beam patterns.** A visualization of the learned beam profiles for 7–, 10– and 20–MLA initializations as presented in the Appendix in Figures 5, 6 and 7, respectively. These profiles suggest that the general trend of the beam transformation is towards higher directivity. The wider the initialized beams are (higher MLA rates), the greater is the increase in the directivity, such that for the very wide 20–MLA initialization (as depicted in Figure 4), the beam pattern converges into two splitted narrower beams. The visualization of the beam profiles of the 10–MLT and 10–random initializations, as displayed in the Appendix in Figures 8 and 9, respectively, suggest that there is a trade-off between the directivity of the beam and the field of view it covers. The 10–MLT profile displays a trend towards widening the simultaneously transmitted narrow beams, whereas for the random initialization, some of the beams stays un-directed and some of them approach the MLT pattern.

#### 4. Conclusion and future directions

We have demonstrated, as a proof-of-concept, that jointly learning the transmit patterns with the receive beamforming provides greater improvements to the image quality. It should be mentioned that since the beam patterns trained from the MLA initialization displayed the optimal results, we can assume the models have not reached the globally optimal configuration – otherwise, all patterns would have converged to similar performance. This calls for better optimization techniques which are more robust to initialization in regression problems in general and in imaging in particular. It should be noted that in all the experiments mentioned within this paper, delay-and-sum

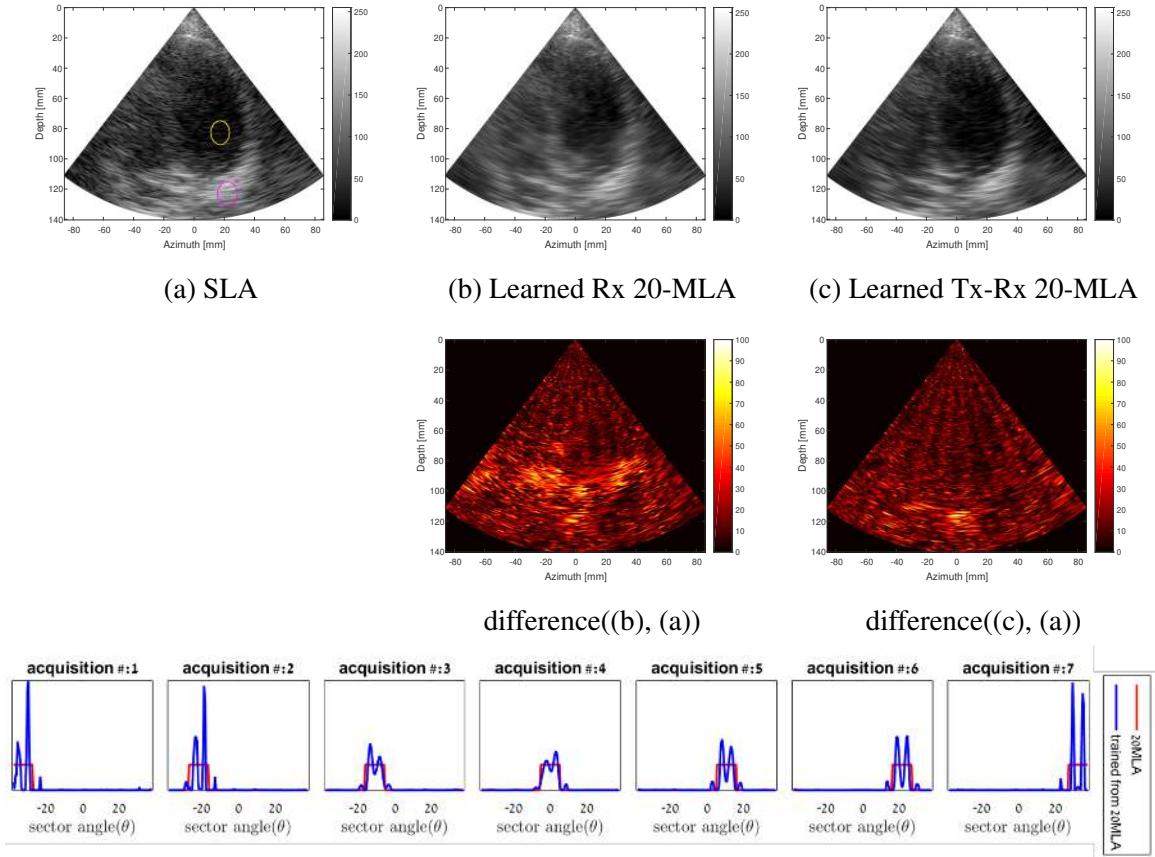


Figure 4: Visual comparison of *Learned-Rx* and *Learned Tx-Rx* settings of 20-MLA on a test frame. The first row depicts (a) the ground truth SLA image, (b) reconstruction obtained from the *Learned Rx* setting and (c) the reconstruction obtained from the *Learned Tx-Rx* setting. The second row depicts the corresponding difference frames (with respect to the SLA image). The bottom row depicts the initial (red) and learned beampatterns (blue) of the 7 acquisitions in the 20-MLA setting.

beamformed SLA was considered as the ground truth reference to the neural network. However, the presented methodology can be simply extended to more sophisticated beamformers such as minimum-variance beamforming by modifying the reference envelope ultrasound image appropriately (Simson et al., 2018), or to other tasks such as estimating the speed-of-sound (Feigin et al., 2018) or the scatterer maps of the tissues (Vedula et al., 2017). It would be particularly interesting to explore such learning-based beam pattern designs to combat the frame-rate vs. resolution tradeoffs in the case of 2D ultrasound probes and to enable efficient computational sonography (Göbl et al., 2018).

An interesting insight observed from the 10-random experiment is that the learned beam profiles perform significantly better than transmitting random undirected beam patterns both with the delay-and-sum and the learned beamformers. This makes us wonder whether transmitting planar

waves is really optimal with a learned receive pipeline. Lastly, in the proposed work, the learned transmit patterns are fixed during post-training. It would be interesting to explore how to design transmit protocols, that are scene or anatomy adaptive, and extend the proposed methodology to the non-linear second-harmonic imaging. We believe that all these directions would initiate a new line of research towards building efficient learning-driven ultrasound imaging.

## Acknowledgments

This research was funded by ERC StG RAPID. We thank Prof. Dan Adam for making his GE machine available to us.

## References

- Seong Ho Chang, SB Park, and Gyu-Hyeong Cho. Phase-error-free quadrature sampling technique in the ultrasonic b-scan imaging system and its application to the synthetic focusing system. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 40(3):216–223, 1993.
- Micha Feigin, Daniel Freedman, and Brian W. Anthony. A Deep Learning Framework for Single-Sided Sound Speed Inversion in Medical Ultrasound. *arXiv e-prints*, art. arXiv:1810.00322, September 2018.
- M. Gasse, F. Millioz, E. Roux, D. Garcia, H. Liebgott, and D. Fribois. High-quality plane wave compounding using convolutional neural networks. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 64(10):1637–1639, Oct 2017. ISSN 0885-3010. doi: 10.1109/TUFFC.2017.2736890.
- Rüdiger Göbl, Diana Mateus, Christoph Hennersperger, Maximilian Baust, and Nassir Navab. Redefining ultrasound compounding: Computational sonography. *CoRR*, abs/1811.01534, 2018. URL <http://arxiv.org/abs/1811.01534>.
- H. Haim, S. Elmalem, R. Giryes, A. M. Bronstein, and E. Marom. Depth estimation from a single image using deep learned phase coded mask. *IEEE Transactions on Computational Imaging*, 4(3):298–310, Sept 2018. ISSN 2333-9403. doi: 10.1109/TCI.2018.2849326.
- Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.
- A. C. Luchies and B. C. Byram. Deep neural networks for ultrasound beamforming. *IEEE Transactions on Medical Imaging*, 37(9):2010–2021, Sept 2018. ISSN 0278-0062. doi: 10.1109/TMI.2018.2809641.
- O. Menashe and A. Bronstein. Real-time compressed imaging of scattering volumes. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1322–1326, Oct 2014. doi: 10.1109/ICIP.2014.7025264.
- Gabriel Montaldo, Mickaël Tanter, Jérémie Bercoff, Nicolas Benech, and Mathias Fink. Coherent plane-wave compounding for very high frame rate ultrasonography and transient elastography. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 56(3):489–506, 2009.

Arun Nair, Trac D. Tran, Austin Reiter, and Muyinatu Lediju Bell. A deep learning based alternative to beamforming ultrasound images. In *ICASSP*, pages 3359–3363, 04 2018. doi: 10.1109/ICASSP.2018.8461575.

Fabrice Prieur, Bastien Dénarié, Andreas Austeng, and Hans Torp. Correspondence-multi-line transmission in medical imaging using the second-harmonic signal. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 60(12):2682–2692, 2013.

Adi Rabinovich, Zvi Friedman, and Arie Feuer. Multi-line acquisition with minimum variance beamforming in medical ultrasound imaging. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 60(12):2521–2531, 2013.

Adi Rabinovich, Arie Feuer, and Zvi Friedman. Multi-line transmission combined with minimum variance beamforming in medical ultrasound imaging. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 62(5):814–827, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Ortal Senouf, Sanketh Vedula, Grigoriy Zurakhov, Alex Bronstein, Michael Zibulevsky, Oleg Michailovich, Dan Adam, and David Blondheim. High frame-rate cardiac ultrasound imaging with deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 126–134. Springer, 2018.

Walter Simson, Nassir Navab, and Guillaume Zahnd. Deepforming: a deep learning strategy for ultrasound beamforming applied to sub-sampled data. *IEEE International Ultrasonics Symposium (IUS)*, 2018.

Nicki Skafte Detlefsen, Oren Freifeld, and Søren Hauberg. Deep diffeomorphic transformer networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4403–4412, 2018.

Sanketh Vedula, Ortal Senouf, Alexander M. Bronstein, Oleg V. Michailovich, and Michael Zibulevsky. Towards ct-quality ultrasound imaging using deep learning. *CoRR*, abs/1710.06304, 2017. URL <http://arxiv.org/abs/1710.06304>.

Sanketh Vedula, Ortal Senouf, Grigoriy Zurakhov, Alex Bronstein, Michael Zibulevsky, Oleg Michailovich, Dan Adam, and Diana Gaitini. High quality ultrasonic multi-line transmission through deep learning. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pages 147–155. Springer, 2018.

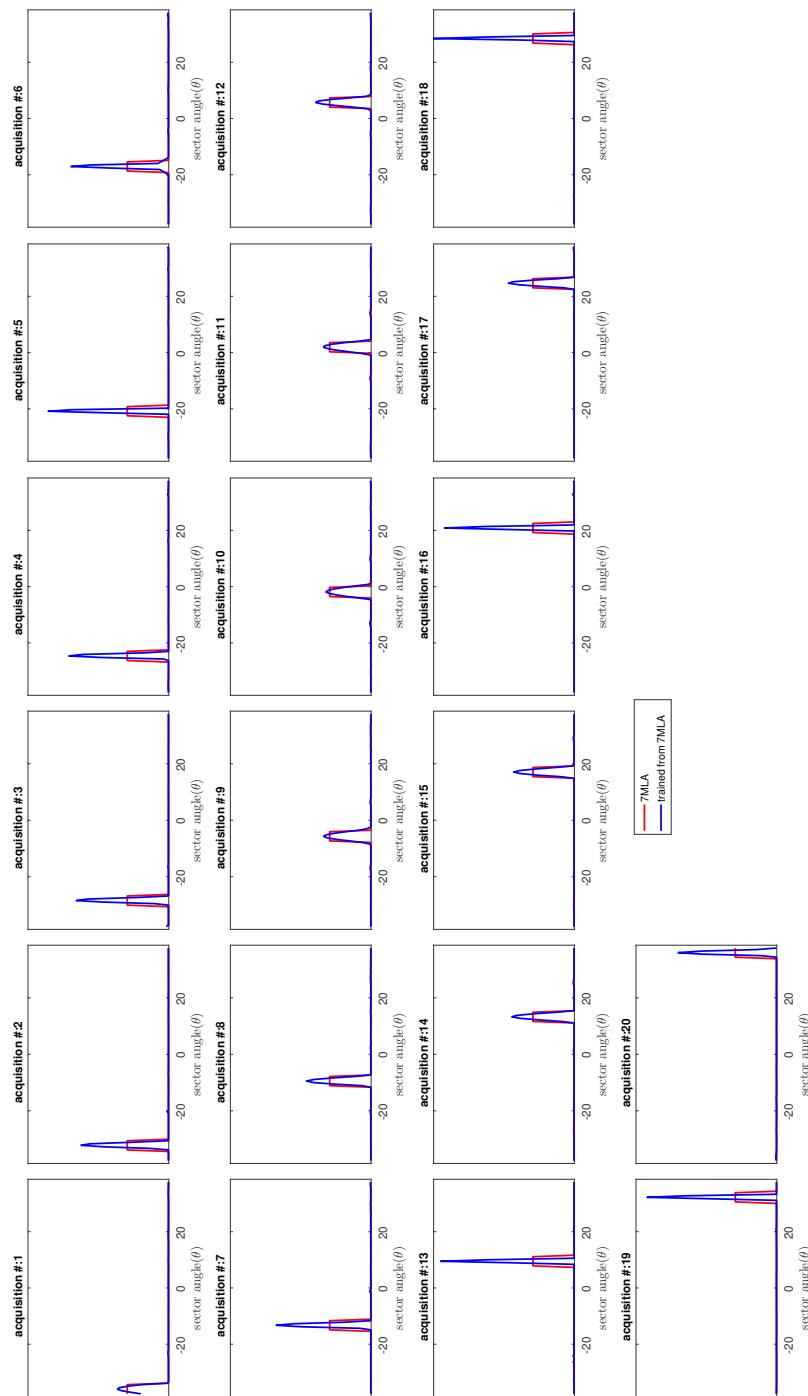


Figure 5: 7-MLA: fixed vs. learned transmission

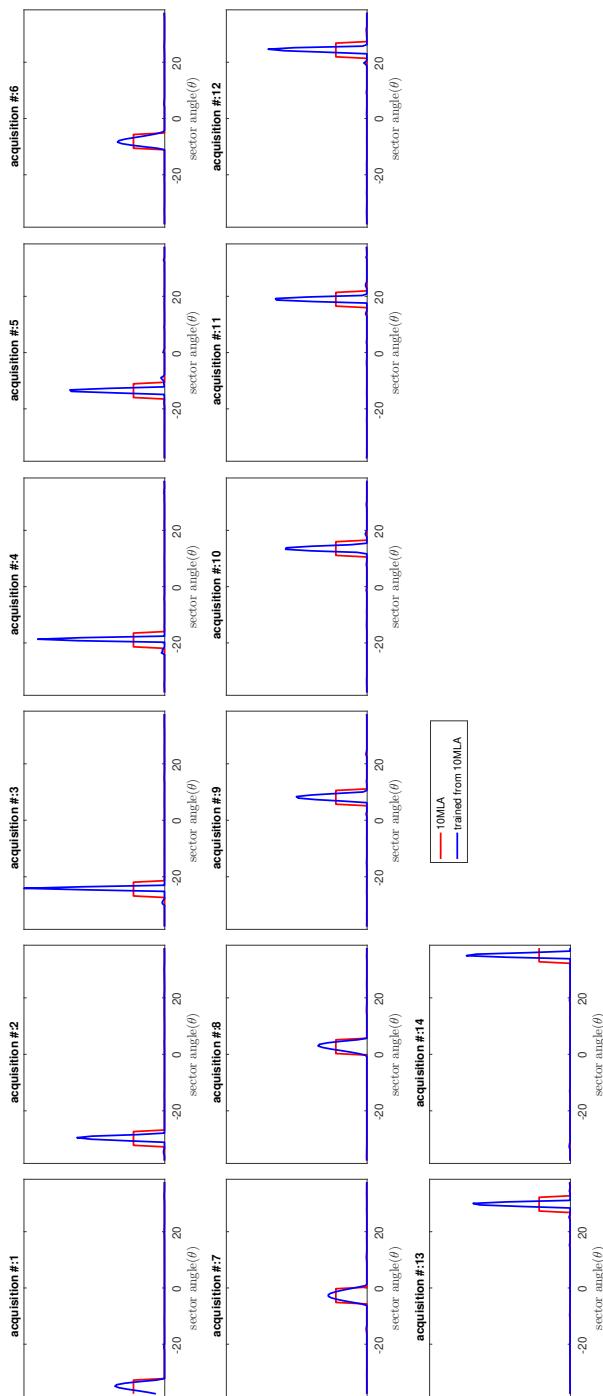


Figure 6: 10-MLA: fixed vs. learned transmission

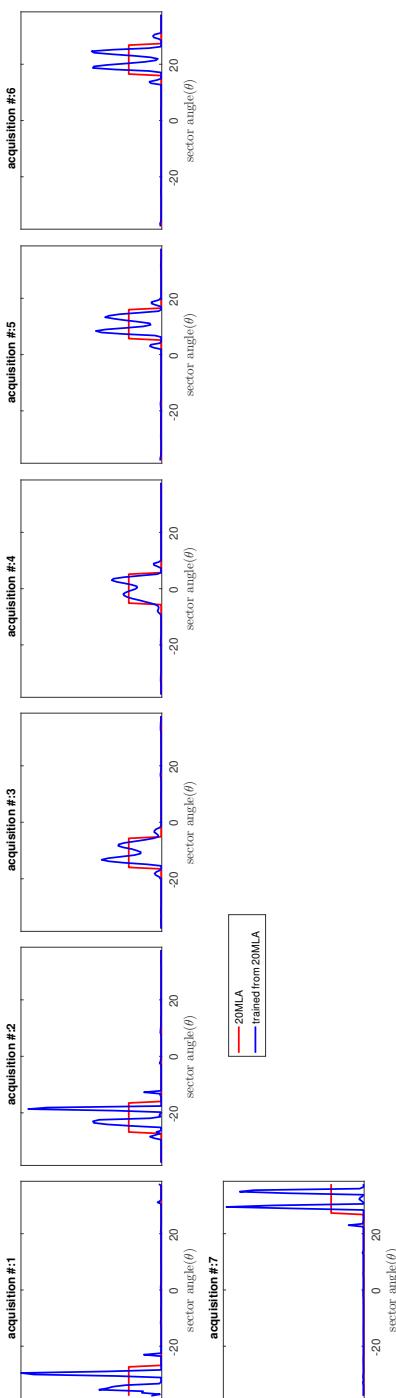


Figure 7: 20-MLA: fixed vs. learned transmission

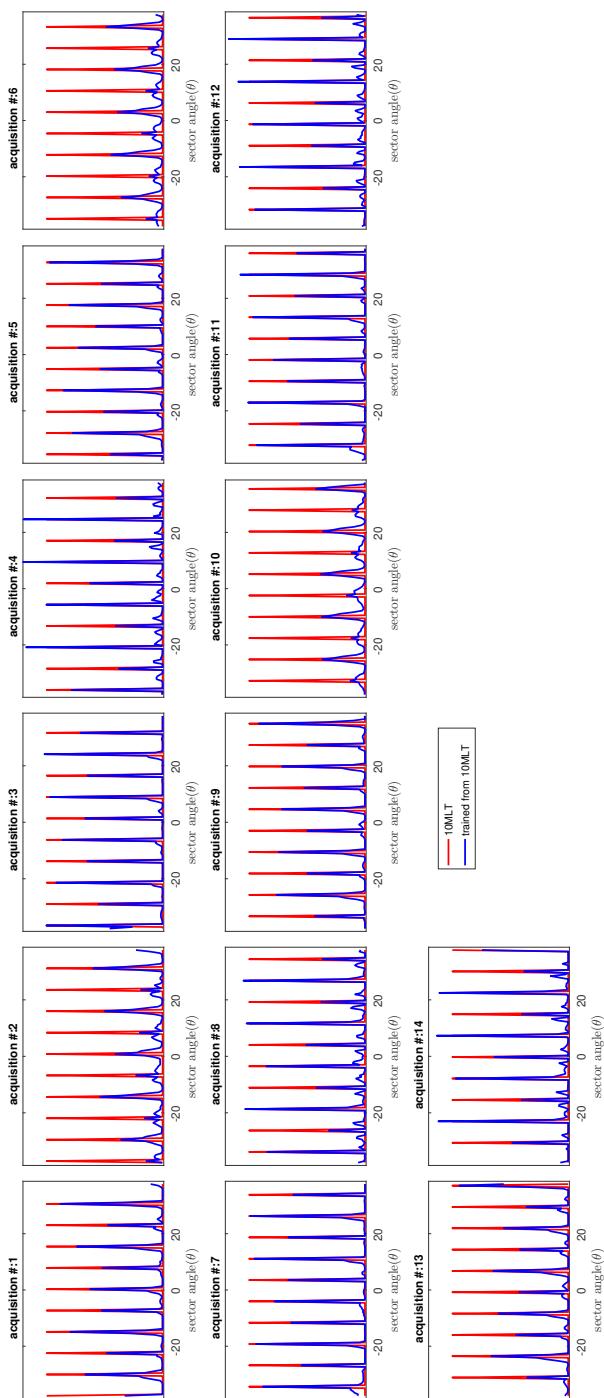


Figure 8: 10-MLT: fixed vs. learned transmission

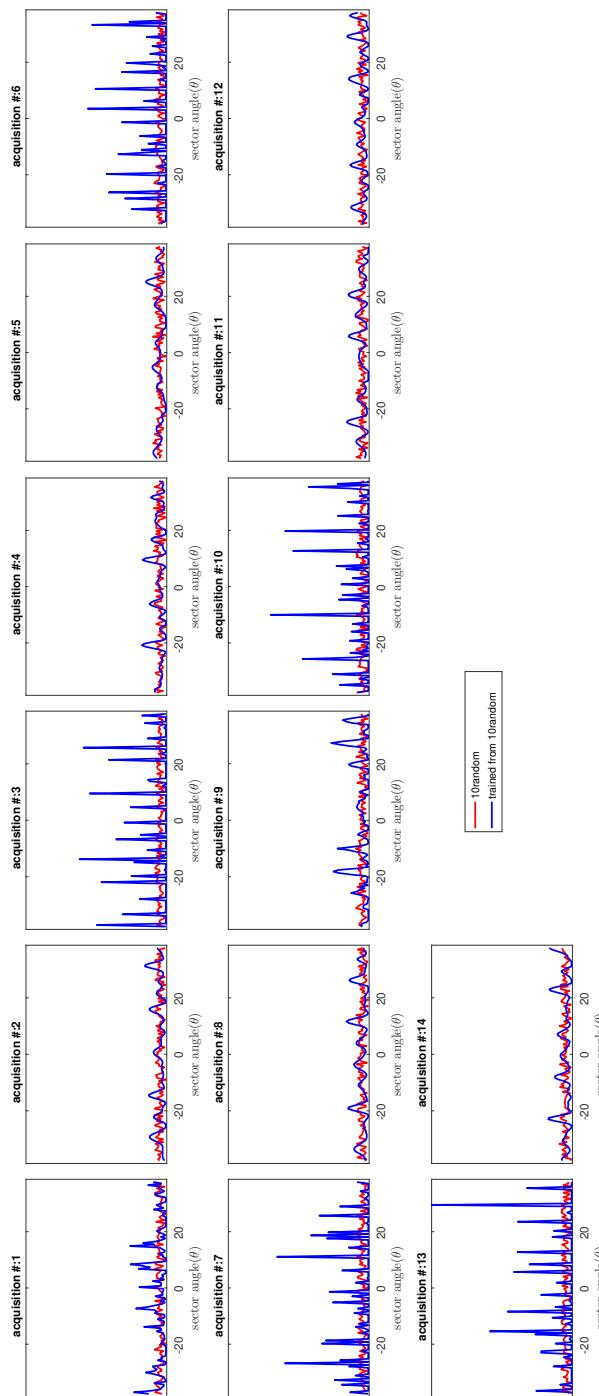
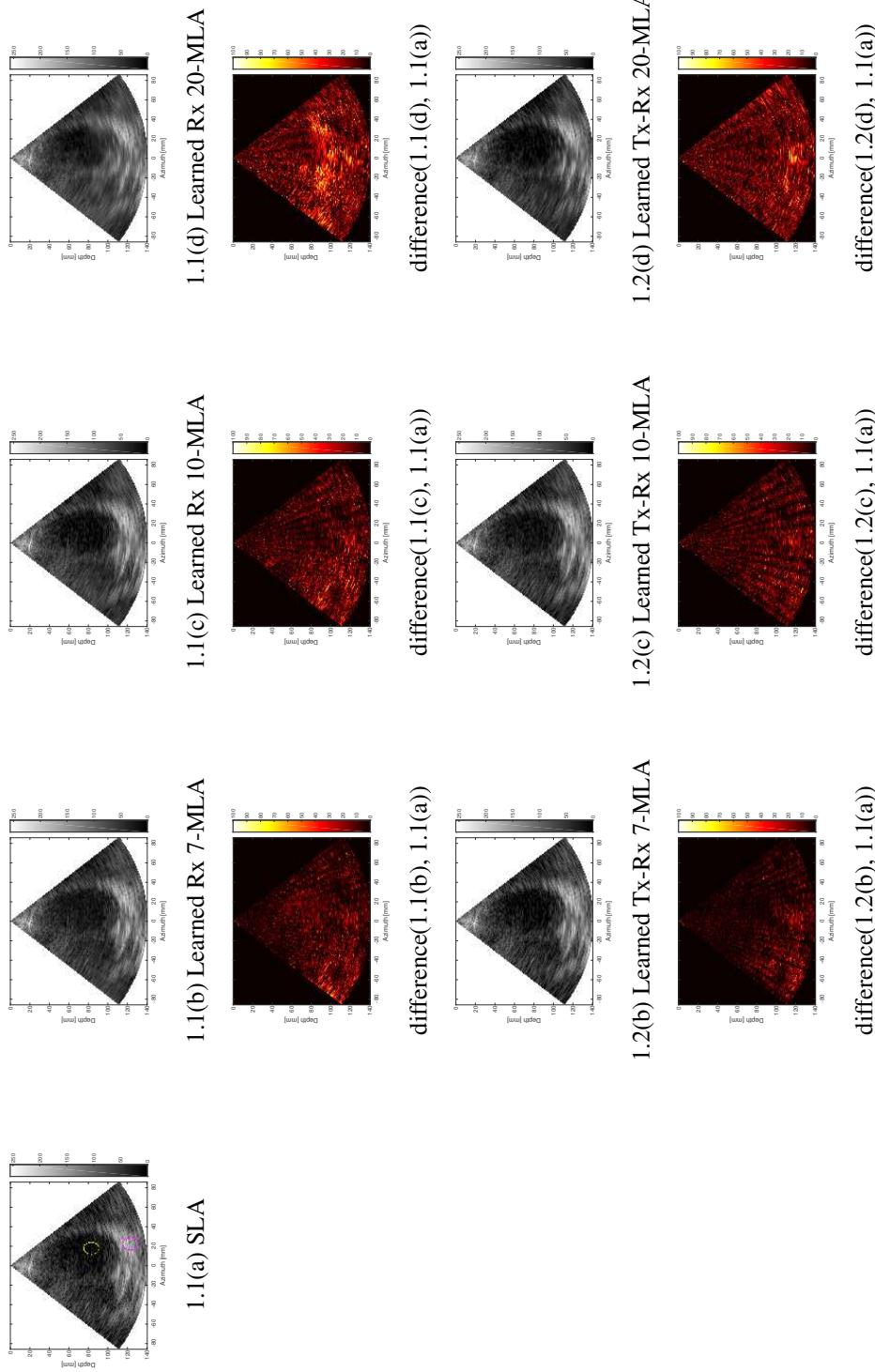
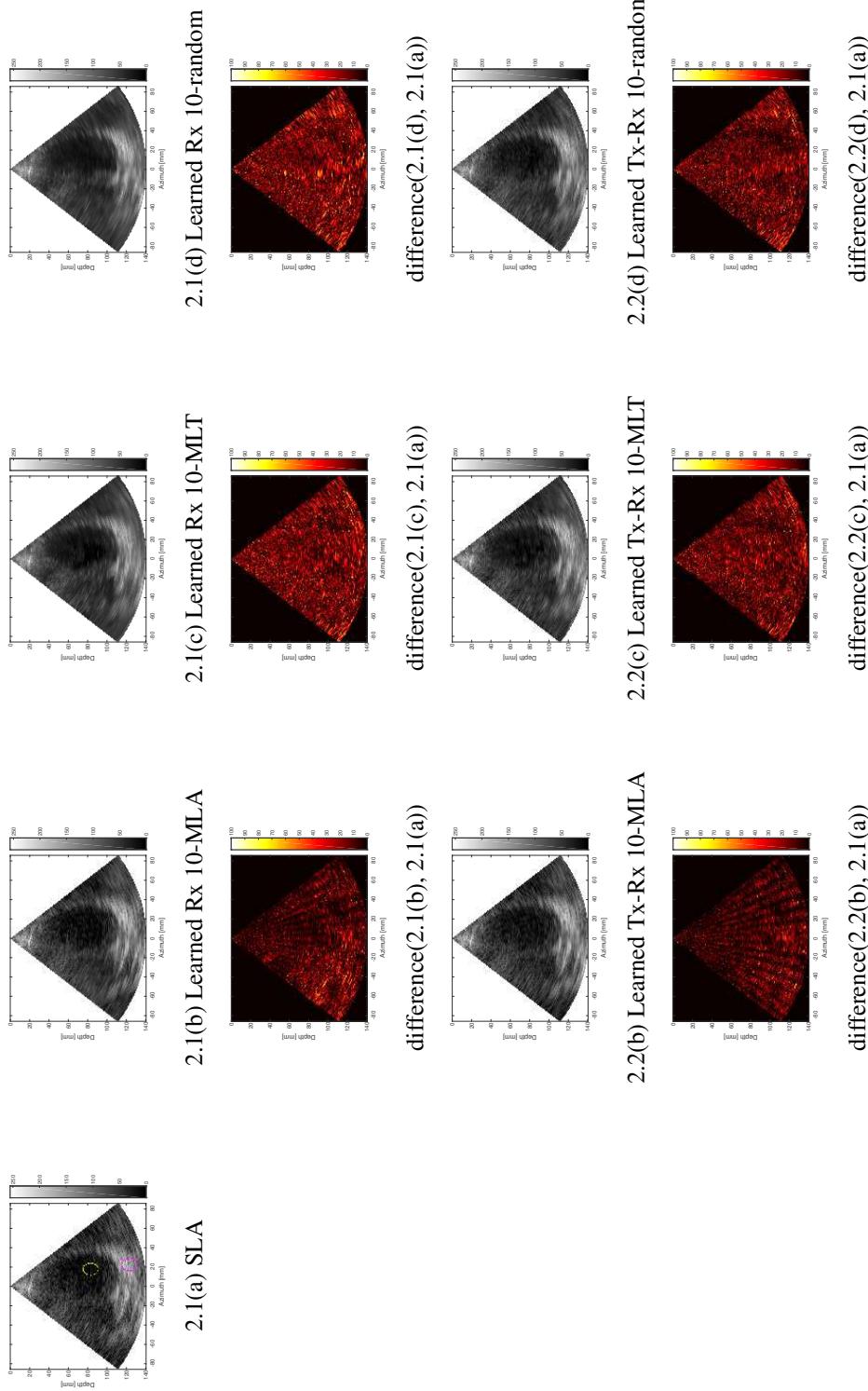


Figure 9: 10-random: fixed vs. learned transmission



**Figure 10: Comparision of Learned Rx vs. Learned Tx-Rx for different initializations.** A test frame from the cineloop comparing the visual results of learned Rx and learned Tx-Rx. The top row depicts the reconstruction obtained from the Learned Rx setting and the third row depicts the reconstruction obtained from the Learned Tx-Rx setting. Even rows depict the difference frames – difference(x, y) indicates the difference between x and y. The difference maps are scaled between [0-100] for better visualization. Digital zoom-in is recommended.



**Figure 11: Comparision of Learned Rx vs. Learned Tx-Rx for different decimation rates.** A test frame from the cineloop comparing the visual results of Learned Rx and Learned Tx-Rx. The top row depicts the reconstruction obtained from the Learned Rx setting and the third row depicts the reconstruction obtained from the Learned Tx-Rx setting. Even rows depict the difference frames – difference(x, y) indicates the difference between x and y. The difference maps are scaled between [0-100] for better visualization. Digital zoom-in is recommended.

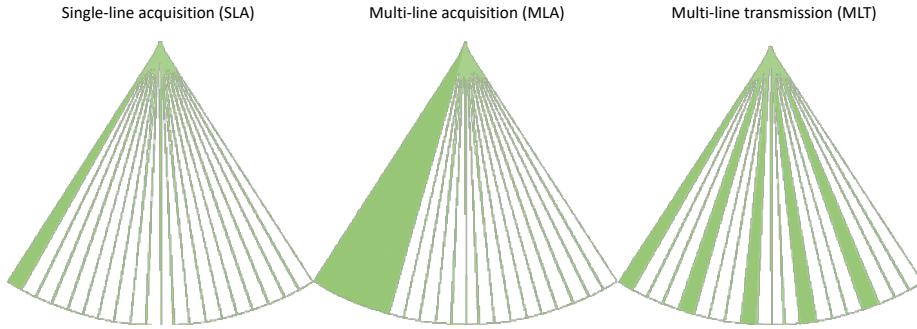


Figure 12: SLA/SLT vs. MLA, MLT

	7-MLA		10-MLA		20-MLA	
	Cr	CNR	Cr	CNR	Cr	CNR
Learned Rx	-30.4463dB	1.3432	-33.2432dB	1.3453	-28.3764dB	1.32
Learned Tx-Rx	-33.2593dB	1.3495	-31.6148dB	1.3891	-32.6599dB	1.3214

Table 3: Comparison of average contrast-to-noise ratio (CNR) and contrast(Cr) measures between different decimation rates of the transmits. Top and bottom rows indicate the results corresponding to learned Rx and learned Tx-Rx experiment settings respectively.CNR and Cr are calculated for the regions marked within yellow and pink circles drawn in Figure 5, 1.1(a).

	10-MLA		10-MLT		10-random	
	Cr	CNR	Cr	CNR	Cr	CNR
Learned Rx	-33.2432dB	1.3453	-28.3089 dB	1.6155	-30.3793dB	1.3452
Learned Tx-Rx	-31.6148dB	1.3891	-28.8051 dB	1.4528	-31.4859dB	1.3418

Table 4: Comparison of average contrast-to-noise ratio (CNR) and contrast(Cr) measures between different initializations of the transmit patterns. Top and bottom rows indicate the results corresponding to learned Rx and learned Tx-Rx experiment settings respectively. CNR and Cr are calculated for the regions marked within yellow and pink circles drawn in Figure 5, 2.1(a).

# Adversarial Pseudo Healthy Synthesis Needs Pathology Factorization

**Tian Xia<sup>1</sup>**

TIAN.XIA@ED.AC.UK

**Agisilaos Chartsias<sup>1</sup>**

AGIS.CHARTSIAS@ED.AC.UK

**Sotirios A. Tsaftaris<sup>1,2</sup>**

S.TSAFTARIS@ED.AC.UK

<sup>1</sup> School of Engineering, University of Edinburgh, West Mains Rd, Edinburgh EH9 3FB, UK

<sup>2</sup> The Alan Turing Institute, London, UK

## Abstract

Pseudo healthy synthesis, i.e. the creation of a subject-specific ‘healthy’ image from a pathological one, could be helpful in tasks such as anomaly detection, understanding changes induced by pathology and disease or even as data augmentation. We treat this task as a factor decomposition problem: we aim to separate what appears to be healthy and where disease is (as a map). The two factors are then recombined (by a network) to reconstruct the input disease image. We train our models in an adversarial way using either paired or unpaired settings, where we pair disease images and maps (as segmentation masks) when available. We quantitatively evaluate the quality of pseudo healthy images. We show in a series of experiments, performed in ISLES and BraTS datasets, that our method is better than conditional GAN and CycleGAN, highlighting challenges in using adversarial methods in the image translation task of pseudo healthy image generation.

**Keywords:** pseudo healthy synthesis, GAN, cycle-consistency, factorization

## 1. Introduction

The aim of pseudo healthy synthesis is to synthesize a subject-specific ‘healthy’ image from a pathological one. Generating such images can be valuable both in research and in clinical applications. For example, these images can be used as a means to perform pathology segmentation (Bowles et al., 2017; Ye et al., 2013), detection (Tsunoda et al., 2014), to help with the visual understanding of disease classification networks (Baumgartner et al., 2018) and to aid experts with additional diagnostic information (Sun et al., 2018).

A challenge with pseudo healthy synthesis is the lack of paired pathological and healthy images for training, i.e. we do not have images of the same patient moments before and after pathology has appeared. Thus, methods based on pure supervised learning are not fit for our purpose. While longitudinal observations could perhaps partially alleviate this problem, the time difference between observations is an additional factor that may complicate learning. Thus, it is imperative to overcome this lack of paired data. One approach is to learn distributions that characterize the domains of healthy and pathological images, for example by learning a compact manifold of patch-based dictionaries (Ye et al., 2013; Tsunoda et al., 2014), or alternatively by learning mappings between the two domains with the use of adversarial training (Sun et al., 2018).

We follow a similar approach here but focus on factorizing the pathology. Simple schematic and examples are shown in Figure 1. We aim to separate what appears to be healthy out of a disease image. We let neural networks decompose an input image into a healthy image (one factor) via a generator, and a binary map that aims to localize disease (the other factor) via a segmentor.

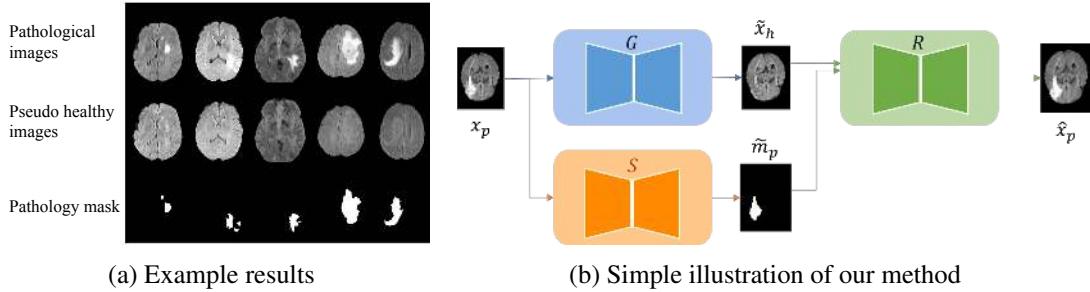


Figure 1: Example results and simple illustration of our method. The three rows of (a) show input pathological images, corresponding pseudo healthy images, and pathology segmentation masks, respectively. Images are taken from the BraTS dataset. In (b) a pseudo healthy image  $\tilde{x}_h$  and a pathology mask  $\tilde{m}_p$  are generated from a pathological image  $x_p$ , and then finally a reconstructed image  $\hat{x}_p$  is generated from  $\tilde{x}_h$  and  $\tilde{m}_p$ .

These two factors are then composed together to reconstruct the input via another network. The pathological map is necessary as a factor to solve the one-to-many problem<sup>1</sup> (Chu et al., 2017): the healthy image must by definition contain ‘less information’ than the disease image.

We can train the segmentor in a supervised way using ‘paired’ pathological images and their corresponding masks. However, since annotations of pathology are not easy to acquire, we also propose an ‘unpaired’ training strategy. We take advantage of several losses including a cycle-consistency loss (Zhu et al., 2017), but use a modified second cycle where we enforce healthy-to-healthy image translation to approach the identity. Finally, since most pseudo healthy methods focus on applications of the synthetic data, results are either evaluated qualitatively or by demonstrating improvements on downstream tasks. A direct quantitative evaluation of the quality of pseudo healthy images has been largely ignored. In this paper, we propose two numerical evaluation metrics for characterizing the ‘*healthiness*’ (i.e. how close to being healthy) and ‘*identity*’ (i.e. how close to corresponding to the input identity) of synthetic results.

Our **contributions** in this work are three-fold:

1. We propose a 2D method that factorizes anatomical and pathological information.
2. We consider two training settings: (a) *paired*: when we have paired images and ground-truth pathology masks; (b) *unpaired*: when such pairs are not available.
3. We propose numerical evaluation metrics to explicitly evaluate the quality of pseudo healthy synthesized images, and compare our method with conditional GAN (Mirza and Osindero, 2014) and CycleGAN (Zhu et al., 2017) on ISLES and BraTS datasets.

## 2. Related work on pseudo healthy synthesis

Medical image synthesis is an active research topic in medical image analysis (Frangi et al., 2018) with an active community and dedicated workshops (e.g. the SASHIMI MICCAI series). For brevity

1. There could be many disease images that could originate from the same healthy image, e.g. consider the simple setting of a lesion in many different locations on the same brain.

here we focus on methods related to pseudo healthy image synthesis with adversarial mechanisms. Image synthesis (translation) can be solved by a conditional GAN that learns a mapping between image domains (e.g. A to B). However, preservation of ‘identity’ is not guaranteed: there are no explicit costs to enforce that an image from domain A to be translated to the same image in domain B. CycleGAN uses a cycle-consistency loss to promote identity and has been profoundly adopted in medical image analysis (Huo et al., 2018; Zhang et al., 2018; Wolterink et al., 2017; Chartsias et al., 2017; Wang et al., 2018).

Baumgartner et al. (2018) used Wasserstein GAN (Arjovsky et al., 2017) to generate disease effect maps, and used these maps to synthesize pathological images. Andermatt et al. (2018) combined the idea of Baumgartner et al. (2018) with CycleGAN to perform pseudo healthy synthesis for pathology detection. Yang et al. (2016) used a Variational Auto-encoder to learn a mapping from pathological images to quasi-normal (pseudo healthy) images to improve atlas-to-image registration accuracy with large pathologies. Schlegl et al. (2017) and Chen and Konukoglu (2018) trained adversarial auto-encoder networks only on normal data, then used the trained model to synthesize normal data from abnormal data as a way of detecting the anomaly. Sun et al. (2018) proposed a CycleGAN-based method to perform pseudo healthy synthesis treating ‘pathological’ and ‘healthy’ as two domains.

The majority of these works use pseudo healthy images to achieve improvements in downstream tasks. While the performance on such downstream tasks relies on pseudo healthy image quality, it is not explicitly evaluated. Herein, we pay particular attention to consistently evaluate how ‘healthy’ the synthetic images look, and whether they correspond to the same ‘identity’ of the input. All methods rely on some form of adversarial training to approximate a distribution. However, as we will detail below, when one of the domains has less information the one-to-many problem can appear and CycleGAN may collapse. Our method treats pathology as a ‘residual’ factor: it factorizes anatomical and pathological information using adversarial and cycle-consistent losses to bypass the one-to-many problem.

### 3. Methodology

#### 3.1. Problem overview

We denote a pathological image as  $x_p$  and a healthy image as  $x_h$ , drawn from  $\mathcal{P}$  and  $\mathcal{H}$  distributions, respectively, i.e  $x_p \sim \mathcal{P}$  and  $x_h \sim \mathcal{H}$ . Our task is to generate a pseudo healthy image  $\tilde{x}_h$  for a sample  $x_p$ , such that  $\tilde{x}_h$  lies in the distribution of healthy images, i.e.  $\tilde{x}_h \sim \mathcal{H}$ . In the meantime, we also want the generated image  $\tilde{x}_h$  to maintain the identity of the original image  $x_p$ , i.e. to come from the same subject as  $x_p$ . Therefore, pseudo healthy synthesis can be formulated as two major objectives: *remove* the disease of pathological images, and *Maintain* the identity and realism as good as possible.

#### 3.2. The one-to-many problem: motivation for factorization

CycleGAN has to somehow invent (or hide) information when one domain contains less information than the other. In our case domain  $\mathcal{P}$  does contain disease information that should not be present in  $\mathcal{H}$ , which leads to failure cases as shown in Figure 2. When CycleGAN cannot invent information, Chu et al. (2017) in fact showed that it hides information within an image to be able to solve the one-to-many mapping. Recently, several papers (Chartsias et al., 2018; Almahairi et al., 2018;

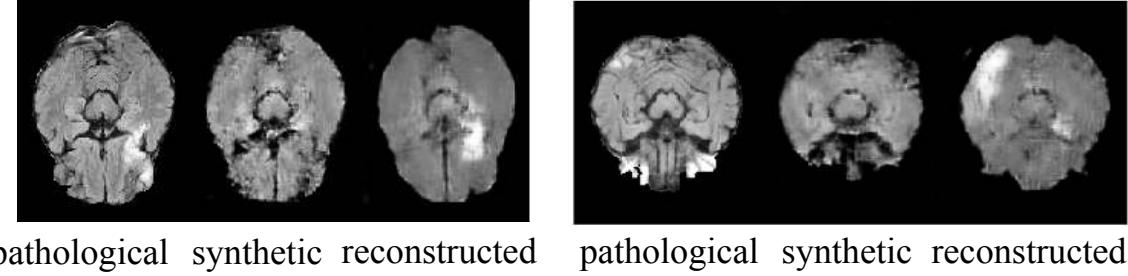


Figure 2: CycleGAN failure cases caused by the one-to-many problem. Each subfigure from left to right shows the input, the pseudo healthy and the input reconstruction. The lesion location in the reconstruction differs from the original one, since an accurate pseudo healthy image has no information to guide the reconstruction process. Images taken from ISLES.

Huang et al., 2018; Lee et al., 2018; Esser et al., 2018) have shown that one needs to provide auxiliary information in the form of a style or modality specific code (actually a vector) to guide the translation and allow many-to-many mappings. Our paper does follow this practice, but instead of providing a vector we consider the auxiliary information of where the disease could be, such that the decoder does *not* have to invent where things should go, and conversely the encoder does *not* have to hide information. We thus achieve that pseudo healthy images are of high quality, correspond to the identity of the same input, and also produce realistic disease maps.

### 3.3. Proposed approach

A schematic of our proposed method is illustrated in Figure 3. Recall that our task is to transform an input pathological image  $x_p$  to a disease-free image  $\tilde{x}_h$  whilst maintaining the identity of  $x_p$ . Towards this goal, our method uses the cycle-consistency losses and treats ‘pathological’ and ‘healthy’ as two image domains. To solve the one-to-many mapping problem, we estimate a disease map from a pathological image using a segmentation network, and then use the map to provide information about disease location. Specifically, there are three main components: ‘ $G$ ’ the ‘generator’; ‘ $S$ ’ the ‘segmentor’; and ‘ $R$ ’ the ‘reconstructor’ trained using two cycles: *Cycle P-H* and *Cycle H-H*.

*Cycle P-H*, we perform pseudo healthy synthesis, where ‘ $G$ ’ takes a pathological image  $x_p$  as input and outputs a ‘healthy’ looking image  $\tilde{x}_h$ :  $\tilde{x}_h = G(x_p)$ . ‘ $S$ ’ takes  $x_p$  as input and outputs a pathology map  $\tilde{m}_p$ :  $\tilde{m}_p = S(x_p)$ . ‘ $R$ ’ takes the synthesized ‘healthy’ image  $\tilde{x}_h$  and the segmented mask  $\tilde{m}_p$  as input and outputs a ‘pathological’ image  $\hat{x}_p$ :  $\hat{x}_p = R(\tilde{x}_h, \tilde{m}_p) = R(G(x_p), S(x_p))$ .

*Cycle H-H* utilizes healthy images and stabilizes the training. It starts with a healthy image  $x_h$  and a null ‘healthy’ mask  $m_h$ . First, ‘ $R$ ’ generates a fake ‘healthy’ image  $\tilde{x}_h$ :  $\tilde{x}_h = R(x_h, m_h)$ , which is then segmented into a healthy mask  $\hat{m}_h$ :  $\hat{m}_h = S(\tilde{x}_h)$  and transformed to a reconstructed healthy image  $\hat{x}_h$ :  $\hat{x}_h = G(\tilde{x}_h)$ .

There are several reasons why we design *Cycle H-H* in such a way. First, a pathology mask for a real healthy image is, by definition, a black mask. Second, we want to prevent the reconstructor ‘ $R$ ’ from inventing pathology when the input disease map is black. Third, we want to guide the generator ‘ $G$ ’ and segmentor ‘ $S$ ’ to preserve identity when the input (to both) is a ‘healthy’ image, such that the synthesized ‘healthy’ image is as similar to the input ‘healthy’ image as possible.

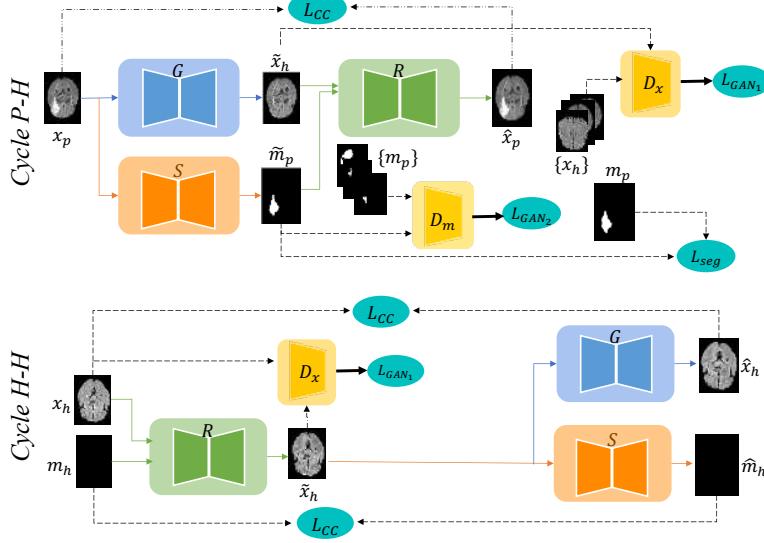


Figure 3: Training flowchart. *Cycle P-H* is the translation path from ‘pathological’ to ‘healthy’ and then back to ‘pathological’; *Cycle H-H* is the path from a healthy image and a black mask to a fake healthy image, then back to the reconstructed image and mask.

Similarly, when the input to the segmentor ‘S’ is a ‘healthy’ image, it should output a ‘healthy’ (no disease) map, i.e. a black mask.

### 3.4. Losses

The training losses are  $\mathcal{L}_{CC}$ ,  $\mathcal{L}_{GAN_1}$  and  $\mathcal{L}_{Seg}$  and  $\mathcal{L}_{GAN_2}$ .

$\mathcal{L}_{CC}$  is the cycle-consistency loss:

$$\begin{aligned} \mathcal{L}_{CC} = & \mathbb{E}_{x_p \sim \mathcal{P}} [\|R(G(x_p), S(x_p)) - x_p\|_1] \\ & + \mathbb{E}_{x_h \sim \mathcal{H}, m_h \sim \mathcal{M}} [\|G(R(x_h, m_h)) - x_h\|_1] + \mathbb{E}_{x_h \sim \mathcal{H}, m_h \sim \mathcal{M}} [\|S(R(x_h, m_h)) - m_h\|_1], \end{aligned}$$

where the first term is defined in *Cycle P-H* and the last two terms are defined in *Cycle H-H*. Note that the third term uses *Mean Average Error* instead of *Dice*, because if the target mask is black, then given any result mask, *Dice loss* will always produce 1.

$\mathcal{L}_{GAN_1}$  is the least squares discriminator loss over synthetic images (Mao et al., 2017):

$$\begin{aligned} \mathcal{L}_{GAN_1} = & \max_{D_x} \min_{G} \frac{1}{2} \mathbb{E}_{x_p \sim \mathcal{P}} [\|D_x(G(x_p)) - 1\|_2] + \max_{D_x} \frac{1}{2} \mathbb{E}_{x_h \sim \mathcal{H}} [\|D_x(x_h)\|_2] \\ & + \max_{D_x} \min_{R} \frac{1}{2} \mathbb{E}_{x_h \sim \mathcal{H}, m_h \sim \mathcal{M}} [\|D_x(R(x_h, m_h)) - 1\|_2] + \max_{D_x} \frac{1}{2} \mathbb{E}_{x_h \sim \mathcal{H}} [\|D_x(x_h)\|_2], \end{aligned}$$

where the first two terms correspond to *Cycle P-H* and the last two for *Cycle H-H*.

To train ‘S’, we use two different training settings whether we have *paired* or *unpaired* data, and use a supervised or a GAN loss, respectively.

In the *paired* setting, we use manually annotated pathology masks corresponding to pathological images in  $\mathcal{L}_{\text{Seg}} = \mathbb{E}_{x_p \sim \mathcal{P}, m_p \sim \mathcal{P}_m} [\text{Dice}(S(x_p) - m_p)]$ , with a differentiable Dice (Milletari et al., 2016) loss.

In the *unpaired* setting, since pathological images lack paired annotations, we replace  $\mathcal{L}_{\text{Seg}}$  with a discriminator  $D_m$  which classifies real pathology masks from inferred masks:

$$\mathcal{L}_{\text{GAN}_2} = \max_{D_m} \min_S \frac{1}{2} \mathbb{E}_{x_p \sim \mathcal{P}} [\|D_m(S(x_p)) - 1\|_2] + \max_{D_m} \frac{1}{2} \mathbb{E}_{m_p \sim \mathcal{P}_m} [\|D_m(m_p)\|_2],$$

where a pathological image  $x_p$  and a mask  $m_p$  come from different volumes.

## 4. Experiments

### 4.1. Experimental settings

**Dataset and preprocessing:** We demonstrate our method on two datasets. We use the FLAIR data of the *Ischemic Lesion Segmentation* (ISLES) 2015 dataset (Maier et al., 2017), which contains images of 28 volumes that are skull stripped and re-sampled to an isotropic spacing of  $1mm^3$  (SISS) resp. We also use FLAIR data from MRI scans of glioblastoma (GBM/HGG), made available in the *Brain Tumour Segmentation* (BraTS) 2018 (Menze et al., 2015) challenge. The BraTS data contain images of 79 volumes that are skull-striped, and interpolated to  $1mm^3$  resolution. Both datasets are released with segmentation masks of the pathological regions. For each dataset, we normalize each volume by clipping the intensities to  $[0, V_{99.5}]$ , where  $V_{99.5}$  is the 99.5% largest pixel value of the corresponding volume, then we normalize the resulting intensities to  $[0, 1]$ . We choose the middle 60 slices from each volume and label a slice as ‘healthy’ if its corresponding pathology mask is black, and as ‘pathological’ otherwise. We divide the datasets into a training and a testing set of 22 and 6 volumes for ISLES, and 50 and 29 volumes for BraTS respectively.

**Training and implementation details:** The method is implemented in Python using Keras (Chollet et al., 2015). The loss function for the paired data option is defined as  $L_{\text{paired}} = \lambda_1 \mathcal{L}_{\text{CC}} + \lambda_2 \mathcal{L}_{\text{GAN}_1} + \lambda_3 \mathcal{L}_{\text{Seg}}$ , where  $\lambda_1 = 10$ ,  $\lambda_2 = 1$ , and  $\lambda_3 = 10$  (same values as Chartsias et al. (2018)). The loss function for the unpaired data option is defined as  $L_{\text{unpaired}} = \lambda_1 \mathcal{L}_{\text{CC}} + \lambda_2 \mathcal{L}_{\text{GAN}_1} + \lambda_3 \mathcal{L}_{\text{GAN}_2}$ , where  $\lambda_1 = 10$ ,  $\lambda_2 = 2$ , and  $\lambda_3 = 10$  ( $\lambda_2$  has been increased to focus the attention on synthesis). Architecture details are in the Appendix.

**Baselines:** We consider two pseudo healthy synthesis baselines for comparison: a *conditional GAN* (Mirza and Osindero, 2014) (that is deterministic and is conditioned on an image) that consists of a pseudo healthy generator, trained with unpaired data and an adversarial loss against a discriminator that classifies real and fake healthy images; and a *CycleGAN* which considers two domains for healthy and unhealthy and is trained as in Zhu et al. (2017) to learn a domain translation using unpaired data.

### 4.2. Evaluation metrics

We propose, and use, numerical evaluation metrics to quantitatively evaluate the synthesized pseudo healthy images in terms of ‘*healthiness*’ and ‘*identity*’ i.e. how healthy do they look and how close to the input they are (as a proxy to identity).

‘*Healthiness*’ is not easy to directly measure since we do not have ground-truth pseudo healthy images. However, given a pathology segmentor applied on a pseudo healthy synthetic image, we can

measure the size of the segmented pathology as a proxy. To this end, we first train a segmentor to predict disease from pathological images, and then use the pre-trained segmentor to predict disease masks of synthetic pseudo healthy images and check how large the predicted disease areas are. Formally, ‘healthiness’ can be defined as:

$$h = 1 - \frac{\mathbb{E}_{\hat{x}_h \sim \mathcal{H}}[N(f_{\text{pre}}(\hat{x}_h))]}{\mathbb{E}_{m_p \sim \mathcal{P}_m}[N(m_p)]} = 1 - \frac{\mathbb{E}_{x_p \sim \mathcal{P}}[N(f_{\text{pre}}(G(x_p)))]}{\mathbb{E}_{m_p \sim \mathcal{P}_m}[N(m_p)]},$$

where  $f_{\text{pre}}$  is the pre-trained segmentor whose output is a pathology mask, and  $N(m)$  is the number of pixels which are labeled as pathology in the mask  $m$ . We normalize by the average size of all ground-truth pathological masks. Then we subtract the term from 1, such that  $h$  increases when the images have smaller pathology.

‘Identity’ is measured using a masked *Multi-Scale Structural Similarity Index* (MS-SSIM) with window width 11, defined as  $\text{MS-SSIM}[(1 - m_p) \odot \hat{x}_h, (1 - m_p) \odot x_p]$ . This metric is based on the assumption that a pathological image and its corresponding pseudo healthy image should look the same in regions not affected by pathology.

#### 4.3. Experiments on ISLES and BraTS datasets

We train our proposed method in both *paired* and *unpaired* settings on ISLES and BraTS datasets, and compare with the baselines of Section 4.1. Some results can be seen in Figure 4, where we observe that all synthetic images visually appear to be healthy. However, the pseudo healthy images generated by *conditional GAN* are blurry and to some degree different from the original samples, i.e. the lateral ventricles (cavities in the middle) change: a manifestation of loss of ‘identity’. Similarly, we observe changes of lateral ventricles in the synthetic images generated by *CycleGAN*. These changes are probably due to the fact that *CycleGAN* needs to hide information to reconstruct the input images. We also observe that our methods preserve more details of the original samples. Together, these observations imply that our proposed methods maintain better ‘identity’ than the baselines.

We also use the proposed evaluation metrics to measure the quality of synthetic images generated by our method and baselines, respectively. The numerical results are shown in Table 1. We can see that our proposed method (paired) when trained using pathological image and mask pairs achieves the best results, followed by our proposed method (unpaired). Both *paired* and *unpaired* versions outperform conditional GAN and CycleGAN in both the BraTS and ISLES datasets. The improvements of our method are due to the factorization of pathology, which ensures maintaining information of the pathology during the pseudo healthy synthesis such that the synthetic images do not need to hide information.

### 5. Conclusion

In this paper, we propose an adversarial network for pseudo healthy synthesis with factorization of pathology. Our proposed method is composed of a pseudo healthy synthesizer to generate pseudo healthy images, a segmentor to predict a pathology map, i.e. as a way of factorizing pathology, and a reconstructor to reconstruct the input pathological image conditioned on the map. Our method can be trained in (a) *paired* mode when we have paired pathological images and masks; or (b) *unpaired* mode for when we do not have image and mask pairs. We also propose two numerical evaluation

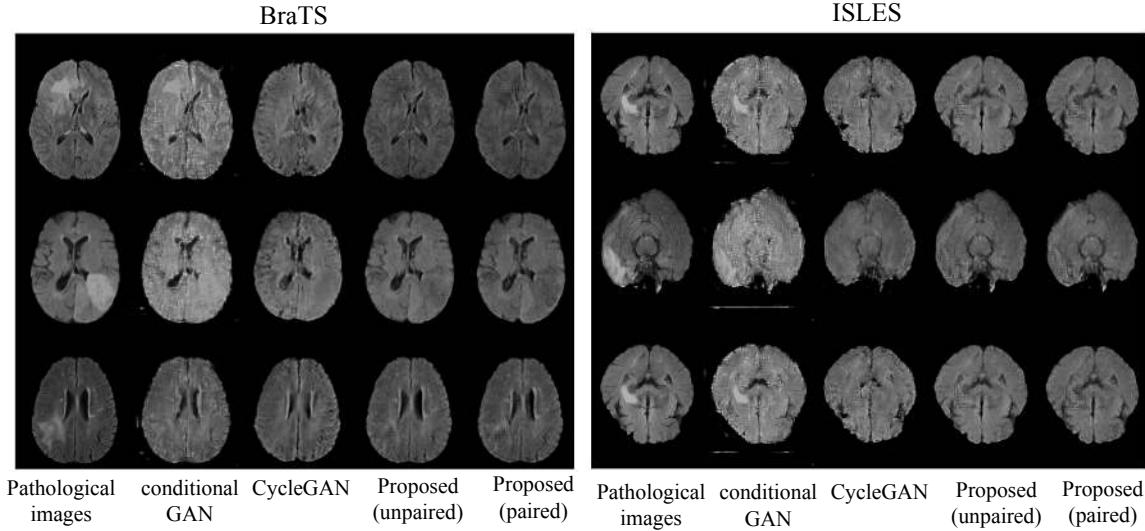


Figure 4: Experimental results for BraTS and ISLES data are shown in the *left* and *right* part respectively. Each part shows three samples (in three rows). The columns from left to right show the ground-truth pathological images, and pseudo healthy images generated by *conditional GAN*, *CycleGAN*, and the two proposed methods, respectively. A larger version of these results are shown in Appendix.

metrics to explicitly measure the quality of the synthesized images. We demonstrate on ISLES and BraTS datasets that our method outperforms the baselines both quantitatively and qualitatively.

Metrics that enforce or even measure identity is a topic of considerable interest in computer vision ([Antipov et al., 2017](#)). Our approach here is simple (essentially measures the fidelity of the reconstructed signal) but it does assume that changes due to disease are only local. This assumption is also adopted by several methods ([Andermatt et al., 2018](#); [Sun et al., 2018](#); [Baumgartner et al.,](#)

Table 1: Evaluation results on BraTS and ISLES of our proposed method trained with and without *pairs*, as well as of the baselines used for comparison. The best mean values for each defined metric (identity, healthiness) are shown in bold. Statistical significant results (5% level), of our methods compared to the best baseline are marked with a star (\*).

Methods	BraTS		ISLES	
	'Identity'	'Healthiness'	'Identity'	'Healthiness'
conditional GAN	$0.74 \pm 0.05$	$0.82 \pm 0.03$	$0.67 \pm 0.02$	$0.86 \pm 0.13$
CycleGAN	$0.80 \pm 0.03$	$0.83 \pm 0.04$	$0.78 \pm 0.02$	$0.85 \pm 0.11$
proposed (unpaired)	$0.83 \pm 0.03$	$0.98 \pm 0.07^*$	$0.82 \pm 0.03$	$0.94 \pm 0.11^*$
proposed (paired)	<b><math>0.88 \pm 0.03^*</math></b>	<b><math>0.99 \pm 0.02^*</math></b>	<b><math>0.93 \pm 0.02^*</math></b>	<b><math>0.98 \pm 0.04^*</math></b>

2018). When disease globally affects an image, new approaches must be devised which is seen as future work.

## Acknowledgments

This work was supported by the University of Edinburgh by a PhD studentship. This work was partially supported by EPSRC (EP/P022928/1) and by The Alan Turing Institute under the EPSRC grant EP/N510129/1. This work was supported in part by the US National Institutes of Health (R01HL136578). We also thank Nvidia for donating a Titan-X GPU.

## References

- Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron C. Courville. Augmented CycleGAN: Learning many-to-many mappings from unpaired data. In *International Conference on Machine Learning*, 2018.
- Simon Andermatt, Antal Horváth, Simon Pezold, and Philippe Cattin. Pathology Segmentation using Distributional Differences to Images of Healthy Origin. *Brain-Lesion workshop (BrainLes). MICCAI*, 2018.
- Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017.
- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.
- Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8309–8319, 2018.
- Christopher Bowles, Chen Qin, Ricardo Guerrero, Roger Gunn, Alexander Hammers, David Alexander Dickie, María Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Brain lesion segmentation through image synthesis and outlier detection. *NeuroImage: Clinical*, 16: 643–658, 2017.
- Agisilaos Chartsias, Thomas Joyce, Rohan Dharmakumar, and Sotirios A Tsaftaris. Adversarial image synthesis for unpaired multi-modal cardiac data. In *International Workshop on Simulation and Synthesis in Medical Imaging (MICCAI)*, pages 3–13. Springer, 2017.
- Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 490–498, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00934-2.

Xiaoran Chen and Ender Konukoglu. Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders. *International Conference on Medical Imaging with Deep Learning*, 2018.

François Chollet et al. Keras. <https://keras.io>, 2015.

Casey Chu, Andrey Zhmoginov, and Mark Sandler. CycleGAN: a Master of Steganography. *NIPS 2017, Workshop on Machine Deception*, 2017.

Patrick Esser, Ekaterina Sutter, and Björn Ommer. A Variational U-Net for Conditional Appearance and Shape Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.

A. F. Frangi, S. A. Tsaftaris, and J. L. Prince. Simulation and Synthesis in Medical Imaging. *IEEE Transactions on Medical Imaging*, 37(3):673–679, March 2018. ISSN 0278-0062. doi: 10.1109/TMI.2018.2800298.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*, volume 11207, pages 179–196. Springer International Publishing, 2018.

Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman. SynSeg-Net: Synthetic Segmentation Without Target Modality Ground Truth. *IEEE Transactions on Medical Imaging*, pages 1–1, 2018. ISSN 0278-0062. doi: 10.1109/TMI.2018.2876633.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse Image-to-Image Translation via Disentangled Representations. In *European Conference on Computer Vision*, volume 11205, pages 36–52. Springer International Publishing, 2018.

Oskar Maier, Bjoern H. Menze, Janina von der Gabeltz, Levin Häni, Matthias P. Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, Daan Christiaens, Francis Dutil, Karl Egger, Chaolu Feng, Ben Glocker, Michael Götz, Tom Haeck, Hanna-Leena Halme, Mohammad Havaei, Khan M. Iftekharuddin, Pierre-Marc Jodoin, Konstantinos Kamnitsas, Elias Kellner, Antti Korvenoja, Hugo Larochelle, Christian Ledig, Jia-Hong Lee, Frederik Maes, Qaiser Mahmood, Klaus H. Maier-Hein, Richard McKinley, John Muschelli, Chris Pal, Linmin Pei, Janaki Raman Rangarajan, Syed M.S. Reza, David Robben, Daniel Rueckert, Eero Salli, Paul Suetens, Ching-Wei Wang, Matthias Wilms, Jan S. Kirschke, Ulrike M. Krämer, Thomas F. Münte, Peter Schramm, Roland Wiest, Heinz Handels, and Mauricio Reyes. "ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI". *Medical Image Analysis*, 35:250 – 269, 2017. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2016.07.009>.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. Weber, T. Arbel, B. B. Avants, N. Ayache,

P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, Oct 2015. ISSN 0278-0062. doi: 10.1109/TMI.2014.2377694.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Fourth International Conference on 3D Vision*, pages 565–571, 2016. doi: 10.1109/3DV.2016.79.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.

Liyan Sun, Jiexiang Wang, Xinghao Ding, Yue Huang, and John Paisley. An Adversarial Learning Approach to Medical Image Synthesis for Lesion Removal. *arXiv preprint arXiv:1810.10850*, 2018.

Yuriko Tsunoda, Masayuki Moribe, Hideaki Orii, Hideaki Kawano, and Hiroshi Maeda. Pseudo-normal image synthesis from chest radiograph database for lung nodule detection. In *Advanced Intelligent Systems*, pages 147–155. Springer, 2014.

Chengjia Wang, Gillian Macnaught, Giorgos Papanastasiou, Tom MacGillivray, and David Newby. Unsupervised learning for cross-domain medical image synthesis using deformation invariant cycle consistency networks. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 52–60. Springer, 2018.

Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Išgum. Deep MR to CT synthesis using unpaired data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 14–23. Springer, 2017.

Xiao Yang, Xu Han, Eunbyung Park, Stephen Aylward, Roland Kwitt, and Marc Niethammer. Registration of pathological images. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 97–107. Springer, 2016.

Dong Hye Ye, Darko Zikic, Ben Glocker, Antonio Criminisi, and Ender Konukoglu. Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 606–613. Springer, 2013.

Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle-and shapeconsistency generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9242–9251, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision*, 2017.

## Appendix A. Architecture details

The detailed architecture of our generator ‘ $G$ ’ is shown in Table 2. IN stands for Instance Normalization. The detailed architecture of our reconstructor ‘ $R$ ’ is shown in Table 3. The detailed architecture of our discriminator ‘ $D_x$ ’ and ‘ $D_m$ ’ is shown in Table 4.

Table 2: Detailed architecture of generator ‘ $G$ ’.

Layer	Input	filter size	stride	IN	activation	Output
conv2d	(208,160,1)	7	1	Yes	ReLu	(208,160,32)
conv2d	(208,160,32)	3	2	Yes	ReLu	(104,80,64)
conv2d	(104,80,64)	3	2	Yes	ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
upsampling2d	(52,40,128)	-	-	-	-	(104, 80, 128)
conv2d	(104, 80, 128)	3	1	Yes	ReLu	(104,80,64)
upsampling2d	(104,80,64)	-	-	-	-	(208, 160, 64)
conv2d	(208, 160, 64)	3	1	Yes	ReLu	(208, 160, 32)
conv2d	(208, 160, 32)	3	1	No	sigmoid	(208, 160, 1)

Table 3: Detailed architecture of reconstructor ‘ $R$ ’.

Layer	Input	filter size	stride	IN	activation	Output
conv2d	(208,160,2)	7	1	Yes	ReLu	(208,160,32)
conv2d	(208,160,32)	3	2	Yes	ReLu	(104,80,64)
conv2d	(104,80,64)	3	2	Yes	ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
residual block	(52,40,128)	3	1	Yes	Leaky ReLu	(52,40,128)
upsampling2d	(52,40,128)	-	-	-	-	(104, 80, 128)
conv2d	(104, 80, 128)	3	1	Yes	ReLu	(104,80,64)
upsampling2d	(104,80,64)	-	-	-	-	(208, 160, 64)
conv2d	(208, 160, 64)	3	1	Yes	ReLu	(208, 160, 32)
conv2d	(208, 160, 32)	3	1	No	sigmoid	(208, 160, 2)

Table 4: Detailed architecture of discriminator ‘ $D_x$ ’ and ‘ $D_m$ ’.

Layer	Input	filter size	stride	IN	activation	Output
conv2d	(208,160,2)	4	2	Yes	Leaky ReLu	(104,80,32)
conv2d	(104,80,32)	4	2	Yes	Leaky ReLu	(52,40,128)
conv2d	(52,40,128)	4	2	Yes	Leaky ReLu	(26,20,256)
conv2d	(26,20,256)	4	2	Yes	Leaky ReLu	(13,10,512)
conv2d	(13,10,512)	4	1	No	sigmoid	(13,10,1)

The detailed architecture of our segmentor ‘ $S$ ’ is a U-Net, and follows the structure of [Ronneberger et al. \(2015\)](#). We change the activation function from ‘ReLU’ to ‘Leaky ReLU’. We also found that using residual connection on each layer slightly improved the results.

The pre-trained segmentor  $f_{pre}$  which is used for evaluation uses the same structure as ‘ $S$ ’. We train the segmentor  $f_{pre}$  on the ISLES and BraTS training datasets (see Section 4.1) respectively, and then use it to evaluate synthetic images generated from samples in ISLES and BraTS testing datasets. The Dice loss of the segmentor on ISLES and BraTS testing datasets are 0.12 and 0.16, respectively.

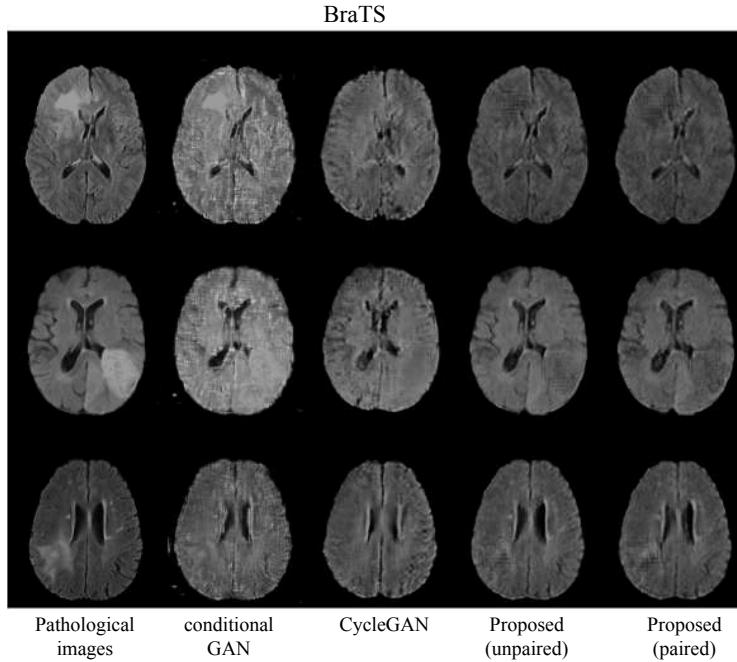


Figure 5: Experimental results for BraTS. The columns from left to right show the ground-truth pathological images, and pseudo healthy images generated by *conditional GAN*, *CycleGAN*, and the two proposed methods, respectively.

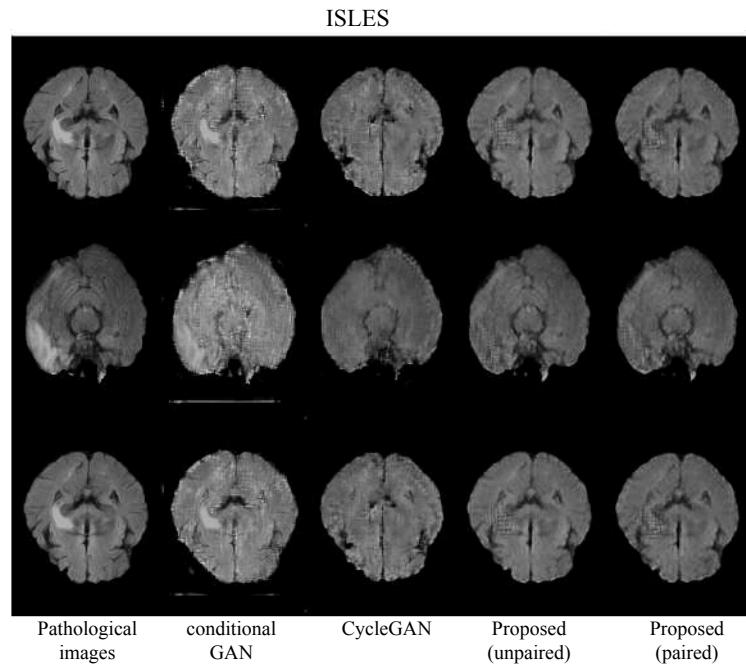


Figure 6: Experimental results for ISLES. The columns from left to right show the ground-truth pathological images, and pseudo healthy images generated by *conditional GAN*, *CycleGAN*, and the two proposed methods, respectively.

# VOCA: Cell Nuclei Detection In Histopathology Images By Vector Oriented Confidence Accumulation

**Chensu Xie<sup>1,2</sup>**

XIC3001@MED.CORNELL.EDU

**Chad M. Vanderbilt<sup>2</sup>**

VANDERBC@MSKCC.ORG

**Anne Grabenstetter<sup>2</sup>**

GRABENSA@MSKCC.ORG

**Thomas J. Fuchs<sup>1,2</sup>**

FUCHST@MSKCC.ORG

<sup>1</sup> *Weill Cornell Graduate School of Medical Sciences, Cornell University, New York, USA*

<sup>2</sup> *Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, USA*

## Abstract

Cell nuclei detection is the basis for many tasks in Computational Pathology ranging from cancer diagnosis to survival analysis. It is a challenging task due to the significant inter/intra-class variation of cellular morphology. The problem is aggravated by the need for additional accurate localization of the nuclei for downstream applications. Most of the existing methods regress the probability of each pixel being a nuclei centroid, while relying on post-processing to implicitly infer the rough location of nuclei centers. To solve this problem we propose a novel multi-task learning framework called vector oriented confidence accumulation (VOCA) based on deep convolutional encoder-decoder. The model learns a confidence score, localization vector and weight of contribution for each pixel. The three tasks are trained concurrently and the confidence of pixels are accumulated according to the localization vectors in detection stage to generate a sparse map that describes accurate and precise cell locations. A detailed comparison to the state-of-the-art based on a publicly available colorectal cancer dataset showed superior detection performance and significantly higher localization accuracy.

## 1. Introduction

Object detection in natural images has been defined as fitting tight bounding boxes around recognized objects. The best examples are the prevailing Fast/Faster-RCNN models (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015) and closely related techniques (Redmon et al., 2016; Liu et al., 2016; He et al., 2017). Cell nuclei detection on histopathology slides requires identification of millions of densely packed small objects per image. This is in contrast to these earlier deep learning works in which usually a few dominant objects are annotated. Due to the several orders of magnitude increase in numbers of objects detected per image, the performance of region proposal based detectors is sub-optimal on cell detection in histology images (Jeong et al., 2017). Further, obtaining annotation of thousands of nuclei bounding boxes is impractical due to the common case of weak nuclei boundaries and high workload of pathologists. To this end, these problems are usually formulated as predicting the  $(x, y)$  coordinates of the objects’ center supervised by point labels (Fuchs et al., 2009).

Most deep learning approaches to cell nuclei detection are based on convolutional neural networks that predict the probability of each pixel being a nucleus centroid (Cireşan et al., 2013; Wang et al.,

2014; Xie et al., 2015b; Chen and Srinivas, 2016; Sirinukunwattana et al., 2016; Zhou et al., 2017; Raza et al., 2018). The final detection of the objects is achieved by identifying the peaks in the probability map using mean shift (Fuchs et al., 2009) or non-maximum suppression (Neubeck and Van Gool, 2006). Fast auto-encoded regression has recently been employed as a technique to explore improved speed and scalability in cell detection over the traditional sliding-window system (Xie et al., 2015a; Chen and Srinivas, 2016; Zhou et al., 2017). Current methods are designed to recognize the cell nuclei and rely on post-processing and *ad hoc* fine-tuning to implicitly infer cell locations, which leads to accumulation of localization error as the number of detected objects gets larger. We must emphasize that while the challenging cell detection is not a clinically useful end as a standalone task, the accurate coordinates of cell nuclei are simply the prerequisite for many downstream applications (e.g. multi-class cell detection for tumor micro-environment analysis, tumor architecture, etc).

To solve this problem, we propose a novel multi-task deep learning method for cell detection. Based on convolutional encoder-decoder, the model concurrently learns 1) binary confidence score, 2) localization vector and 3) weight of contribution for each pixel. In detection stage, the confidence scores are weighted and accumulated to the positions pointed by the localization vectors. We call this method vector oriented confidence accumulation (VOCA). We demonstrate that the three closely correlated but distinct tasks are mutually beneficial when trained as an integrated model (Section 5.1). VOCA explicitly learns the location of nuclei centroid and thus produces profoundly peaked accumulator maps which describe accurate and precise nuclei locations, and enables fast and robust post-processing (Section 5.2). Comparison experiments based on a publicly available colorectal cancer dataset (Sirinukunwattana et al., 2016) shows that our proposed method outperforms the existing methods in terms of F1 score for cell detection, and gives significantly higher nuclei localization accuracy (Section 5.3).

## 2. Related work

Early attempts at cell nuclei detection utilized human expert-designed features describing intensity distribution and morphological patterns (Cosatto et al., 2008; Al-Kofahi et al., 2010; Kuse et al., 2011; Arteta et al., 2012; Ali and Madabhushi, 2012; Veta et al., 2013; Vink et al., 2013). It is notable that many of these works confabulate the related but separate concepts of nuclei detection and segmentation. This confusion is likely because hand-crafted features are often shape oriented. These approaches tend to be brittle due to the significant heterogeneity of histology slides and cellular morphology and require additional engineering and tuning between different datasets.

Recent works employing deep learning for cell nuclei detection have achieved state-of-the-art results. Cireşan et al. (2013) utilized deep neural network to differentiate between mitotic nuclei and background. Cruz-Roa et al. (2013) and Xu et al. (2016) learned unsupervised features via auto-encoders for cell detection, which was extended by Wang et al. (2014) by combining hand-crafted features with deep learning. While object detection at its heart is the combination of object recognition and localization, these works depending on pixel-wise binary classification only considered the first task. Xie et al. (2015b) proposed a structured regression approach to predict the probability of each position being a nucleus centroid. Their regression targets embedded the localization information by formulating the score as a function of the distance ( $d$ ) between each pixel and the nearest

ground truth nucleus. This spirit of integrating the two tasks was also followed by many other works. For example, [Chen and Srinivas \(2016\)](#) labeled pixels for lymphocytes detection by thresholding  $d$ . [Sirinukunwattana et al. \(2016\)](#) proposed a spatially constrained CNN (SC-CNN) regressing to a similar map and published a dataset for nuclei detection on colorectal cancer images. [Zhou et al. \(2017\)](#) developed a sibling fully convolutional network (FCN) architecture for simultaneous cell detection and fine-grained classification. [Raza et al. \(2018\)](#) proposed a framework to deconvolve filter-mapped CNN output for cell detection on lung cancer slides. Considering the variation in nuclei size, [Koohababni et al. \(2018\)](#) formulated each nucleus as a Gaussian peak with a maximum value on its centroid, and directly regress the means and standard deviations with a small image patch as input. [Tofighi et al. \(2018\)](#) utilized additional annotation to combine shape priors with deep features for cell detection. Notably, [Ahmad et al. \(2018\)](#) learned features by correlation filters and achieved state-of-the-art performance for nuclei detection on the previously mentioned colorectal dataset ([Sirinukunwattana et al., 2016](#)) against which several of the above mentioned works were benchmarked. In contrast to these works, VOCA formulates the cell nuclei detection problem as a multi-task approach, which disentangles rather than integrates the objectives, hypothesizing that simpler objectives can potentially improve model training and understanding.

### 3. Method

#### 3.1. Deep multi-task learning

We propose a novel CNN based deep multi-task learning method for cell detection. Each pixel of a training image is scored with 3 tasks. Let  $p_I[i, j]$  be the pixel at coordinate  $(i, j)$  of input image  $I$ , and  $c_I[u, v]$  be the nearest ground truth annotation for a cell nuclei which is at position  $(u, v)$ .  $Conf_I$ ,  $Loc_I$ , and  $Wt_I$  be the target maps of confidence score, localization vector and weight of contribution of image  $I$  respectively. First,

$$Conf_I[i, j] = \begin{cases} 1, & \text{if } \|(u - i, v - j)\|_2 < r \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$r$  is the hyperparameter thresholding the proximity of cells. The confidence score target map indicates whether each pixel should be regarded as a nucleus. The second task

$$Loc_I[i, j] = (u - i, v - j), \quad \text{if } Conf_I[i, j] = 1 \quad (2)$$

is a vector describing the direction and magnitude that  $p_I(i, j)$  needs to move to the location of its assigned ground truth  $c_I(u, v)$ . Note that only pixels labeled as foreground by the confidence map ( $Conf_I[i, j] = 1$ ) are trained with this task. The third task scores  $p_I[i, j]$  as:

$$Wt_I[i, j] = \sum_{c_I[u', v']} \mathbb{1}_{\|(u' - i, v' - j)\|_2 < r}(c_I[u', v']) \quad (3)$$

where  $\mathbb{1}_{\|(u' - i, v' - j)\|_2 < r}(c_I[u', v'])$  is an indicator function of whether a ground truth cell nucleus  $c_I[u', v']$  is within euclidean distance  $r$  to  $p_I[i, j]$ . This task counts the number of cell nuclei that intersect at  $p_I[i, j]$ . Since the pixels lying in the intersection of cells are shared in confidence accumulation (cf. Section 3.3), their contribution should be up-weighted accordingly by  $Wt$ .

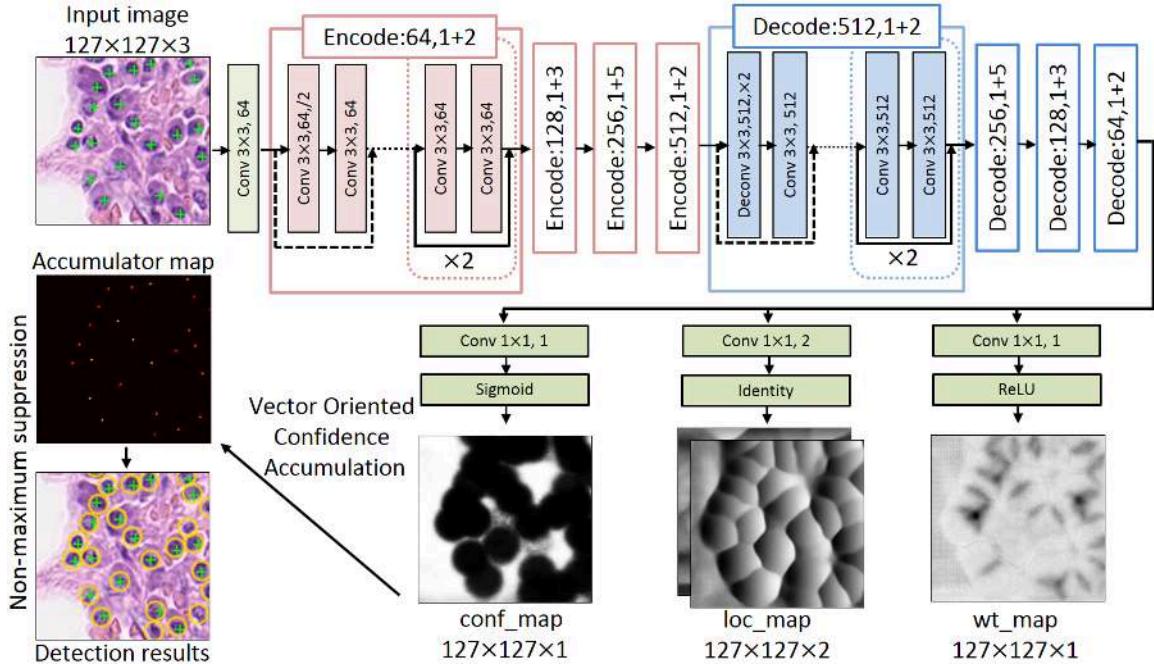


Figure 1: Residual encoder-decoder architecture of our proposed method.

We used binary cross entropy weighted by the inverse of class frequencies as the loss function for confidence score ( $L_{conf}$ ). Inspired by [Girshick \(2015\)](#), we used smooth  $l1$  loss for localization vector and weight of contribution ( $L_{loc}$ ,  $L_{wt}$ ) to avoid gradient explosion. The joint loss function is a linear combination of the three losses:

$$L = L_{conf} + \lambda_1 L_{loc} + \lambda_2 L_{wt} \quad (4)$$

where  $\lambda_1$  and  $\lambda_2$  are parameters weighting the contribution of different tasks. We kept both  $\lambda_1$  and  $\lambda_2$  at 1 in all of our experiments unless discussed (cf. Section 5.1).

### 3.2. Network architecture

Instead of computing a small patch around each pixel in the sliding-window manner, we used an FCN-like structure ([Long et al., 2015](#)) with rich features in the decoding part ([Chen and Srinivas, 2016](#)) to learn the task maps. This design shared convolutional layers and largely reduced the effective input size from the sliding-window approaches. The network abstracts and decodes distinct features for different tasks. The bottom panel of Figure 1 shows the 3 task maps. The confidence score map describes the proximity of nuclei as surrounding disks. The localization vector map is composed of two gradient images zeroed at nuclei position in both  $x$  and  $y$  dimensions. The last map correctly up-weighted the pixels at nuclei intersections. All colors were inverted for improved visualization.

Our proposed model takes input of size  $127 \times 127 \times 3$  and feeds it forward to 4 encoding and 4 decoding blocks followed by 3  $1 \times 1$  conv layers to produce the task maps. We used residual layers for

each block of the encoder-decoder (cf. Figure 1). Rather than max pooling, down/up-sampling was conducted within every block by  $3 \times 3$  conv/deconv layers at stride 2 to retain location information. Changing the receptive field size of the last encoding block by either decreasing or increasing the number of encoding blocks degraded the detection performance in our experiments. We surmise that having a receptive field that is approximately the size of cell nuclei ( $16 \times 16$ ) on cancer slides at  $20 \times$  magnification allows the network to learn higher level semantics useful for the tasks. On top of the last  $1 \times 1$  conv layers, we used sigmoid activation for confidence score maps, since it is stable to our binary cross entropy loss  $L_{conf}$ . Identity function was employed as the activation to account for both negative and positive values of the regression target. For the weight of contribution map we selected ReLU as the activation to learn the positive cell counts.

### 3.3. Vector oriented confidence accumulation

In detection stage, the predicted task maps are combined intuitively to generate an accumulator map (cf. Figure 1). Let  $P$  be a map initialized with zeros. For every coordinate  $(i, j)$ , the localization vector accumulates the weighted confidence score of pixel to the target position :

$$P[i', j'] = P[i', j'] + Wt[\hat{i}, \hat{j}] \times \hat{Conf}[i, j], \text{ where } (i', j') = (i, j) + Loc[\hat{i}, \hat{j}] \quad (5)$$

The confidence accumulation amplifies the stratification between fore-ground and back-ground and produces sparse response, which enhances the speed and robustness of the follow-up non-maximum suppression on  $P$  to output the final detection results.

## 4. Dataset and implementation details

We validated our method on the publicly available colorectal cancer dataset released by [Sirinukunwattana et al. \(2016\)<sup>1</sup>](#). The dataset contains 100 images of size  $500 \times 500$  at  $20 \times$  magnification, which were cropped from 10 whole-slide images of 9 patients with colorectal adenocarcinomas. On these images there are in total 29,747 cell nuclei marked at/around the center. We randomly split the dataset for 2-fold cross validation. The image ids for each subsample is attached in Appendix A.

The network was implemented with PyTorch ([Paszke et al., 2017](#)). Images of size  $127 \times 127$  were further cropped from the dataset by a uniform grid of stride 17 for translational augmentation and to match the model input size. We used batch size 8 and learning rate 0.0005 with a decay factor of 0.1 after every 3 epochs. A momentum of 0.9 was used. Input images were normalized by the mean and standard deviation calculated on the training dataset. For further data augmentation, each image has 50% chance to be flipped horizontally and then 50% chance to be flipped vertically, finally equal chances to be rotated by  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$  and  $270^\circ$  counterclockwise. The model was trained on a single GPU within 4 hours for 10 epochs.

Table 1: Pixel-wise classification accuracy ( $Acc$ ) and localization loss ( $L_{loc}$ ) of training configurations with different combinations of losses.

Configuration	Fold 1		Fold 2	
	$Acc$	$L_{loc}$	$Acc$	$L_{loc}$
Conf	0.879	-	0.882	-
Loc	-	3.969	-	4.077
Conf+Loc	0.886	3.971	0.887	4.071
Conf+Loc+Wt	<b>0.886</b>	<b>3.967</b>	<b>0.887</b>	<b>4.061</b>

## 5. Experiments and discussion

### 5.1. Pixel-wise classification accuracy and localization loss

We first evaluated the effectiveness of multi-task learning. We experimented with different values of the proximity parameter  $r$  in Equation (1) and set it to 12 for all following comparisons as it gave the best F1 score in our cross validation (cf. Section 5.3). A pixel  $p_I[i, j]$  is classified correctly if  $\hat{Conf}[i, j] > 0.5$  and  $Conf[i, j] = 1$ . The pixel-wise classification accuracy ( $Acc$ ) is then defined as the average accuracy of fore-ground and back-ground pixels since we have quite imbalanced sample sizes. As we mentioned before, the localization loss ( $L_{loc}$ ) was calculated as the averaged sum of smooth  $l_1$  losses of both  $x$  and  $y$  dimensions for all pixels. In Table 1 we presented the  $Acc$  and  $L_{loc}$  of different training configurations. Conf+Loc+Wt means that all three losses were trained concurrently. Conf means that only  $L_{conf}$  was used for training. The rest configurations are defined in a similar fashion.

The results imply that the three related tasks are mutually beneficial. Especially the classification accuracy was improved if trained together with localization loss. This improvement (from 0.879 to 0.886 for Fold 1, and from 0.882 to 0.887 for Fold 2) was comparable to other optimization of the pipeline.  $L_{conf}$  and  $L_{wt}$  converges about 3 times faster than  $L_{loc}$  during training. We surmise that regression of localization vector is a more challenging objective therefore contributed more to the learning of common features. We tried various values of  $\lambda_1$  in Equation 4 (while keeping  $\lambda_2$  as 1): 0.1, 1, and 10, but 1 resulted in the best performance. A natural extension of our work would be experimentation with more combinations of the weighting parameters  $\lambda_1$  and  $\lambda_2$ . It is notable that the  $L_{loc}$  almost falls under 4, which is in  $l_1$  form since  $> 1$ . It means that the average localization error on each dimension is only 2 pixels. This observation is consistent with the crisp accumulator maps in Figure 2 and the high localization accuracy shown in Table 2.

### 5.2. Accumulator map and qualitative results

We present in Figure 2 the accumulator maps and qualitative detection results generated by VOCA. For comparison, we also implemented a pixel-wise peak regression model (PR) similar to Xie et al. (2015b). The PR model replaces the multi-task maps of VOCA by a single regression map, in

1. The dataset is available at <https://www2.warwick.ac.uk/fac/sci/dcs/research/tia/data>

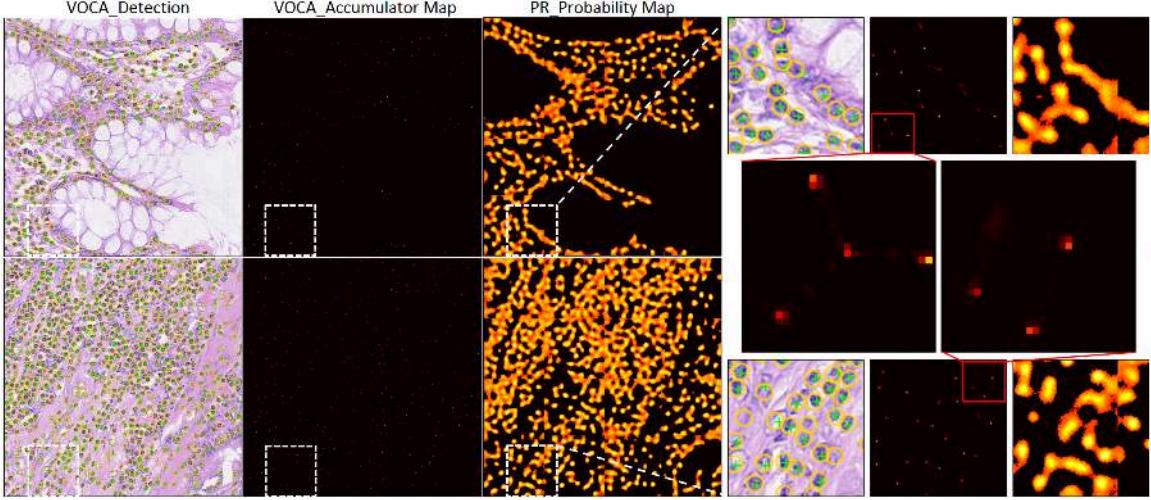


Figure 2: Accumulator maps and cell detection results of VOCA compared to peak regression (PR).  
The figure is best viewed on screen with magnification 400%

which the pixels are scored as  $P_l(i, j) = \begin{cases} \frac{1}{(1+0.8 \times \| (u-i, v-j) \|_2)}, & \text{if } \| (u-i, v-j) \|_2 < 6 \\ 0, & \text{otherwise} \end{cases}$ . It is a representative of several other existing methods (Chen and Srinivas, 2016; Sirinukunwattana et al., 2016; Raza et al., 2018) which also embed recognition and localization to a single map. In detection results (cf. Figure 2 left panel), the yellow circles represent the predicted location and the green crosses are ground truth annotation. Only predictions above the confidence threshold that gives the best F1 score were shown.

As shown in the zoomed-in panels in Figure 2, the predicted confidence scores (cf. *conf\_map* in Figure 1) were accumulated precisely to the target locations. Pixels with high accumulated confidence are within distance of 1 to 2 pixels to the peaks, while the majority of the background becomes zero-valued after confidence “movement”. Post-processing on the clean accumulator maps of VOCA is fast. For example, it speeds up non-maximum suppression whose running time is  $O(\ln(n))$ , where  $n$  is the number of positively valued pixels. In our experiments it took on average 0.2 seconds to process each map of size  $500 \times 500$ , which is about 30 times as fast as on the probability maps produced by PR (cf. Figure 2 mid panel). Besides precision, nuclei localization of VOCA also showed high accuracy as most of the yellow circles (predictions) are rigorously centered at the green crosses (ground truth). The quantitative measurement of the localization accuracy will be presented in Section 5.3.

### 5.3. Quantitative performance and localization accuracy

Non-maximum suppression on the crisp accumulator maps produced by VOCA is not only fast but also robust. A distance threshold of 4 pixels can already suppress most of the non-peak positions. The accumulated scores within 2 pixels of a nucleus coordinate given by non-maximum suppression

Table 2: Comparison of precision, recall, F1 score and localization accuracy

Methods	Precision	Recall	F1 score	Median Distance (Q1, Q3)
LIPSYM	0.725	0.517	0.604	2.236 (1.414, 7.211)
SSAE	0.617	0.644	0.630	4.123 (2.236, 10)
SC-CNN	0.781	0.823	0.802	2.236 (1.414, 5)
SP-CNN	0.803	0.843	0.823	-
MDN	0.788	<b>0.882</b>	0.832	-
SFCN-OPI	0.819	0.874	0.834	-
RBF-CF	0.83	0.86	0.84	-
VOCA-NW	0.814	0.854	0.834	2.0 (1.414, 2.236)
<b>VOCA</b>	<b>0.831</b>	0.863	<b>0.847</b>	<b>2.0 (1.414, 2.236)</b>

were summed as its final score. All scores were normalized to [0, 1] for each image. The predicted coordinates were then assigned to ground truth cell nuclei by Hungarian algorithm (Kuhn, 1955) according to euclidean distance to ensure that at most 1 prediction will be considered true positive for each ground truth. The predictions are regarded as true positive if and only if they are within 6 pixels of their assigned nuclei as suggested by Sirinukunwattana et al. (2016). We plotted precision-recall curves by thresholding the final scores and obtained the optimal F1 score for comparison with the existing methods validated on the same dataset (cf. Table 2). The corresponding precision and recall were also reported.

The first panel of methods (LIPSYM (Kuse et al., 2011), SSAE (Xu et al., 2016), SC-CNN (Sirinukunwattana et al., 2016)) were (re-)validated by Sirinukunwattana et al. (2016) when they published the dataset. The second panel includes the reported results on the same dataset of more recent methods described in Section 2 (SP-CNN (Tofighi et al., 2018), MDN (Koohababni et al., 2018), SFCN-OPI (Zhou et al., 2017), RBF-CF (Ahmad et al., 2018)). VOCA-non-weighted (VOCA-NW) represents our configuration Conf+Loc (cf. Table 1) in which  $W_t$  was not trained and the confidence was thus not weighted for accumulation. “-” means the score is not available from the original paper.

VOCA achieved the best detection performance with F1 score as 0.847. It tends to have higher precision than the other methods at similar recall, which we surmise is caused by its amplification of the stratification between fore-ground and back-ground by confidence accumulation. As  $W_t$  didn’t help the training (cf. Table 1), the improved performance of VOCA over VOCA-NW should come from the compensatory upweighting for pixel sharing during confidence accumulation. Theoretically VOCA-NW gives lower confidence scores for packed cells, since only a portion of the pixels at their intersections (the dark areas in the  $W_t$  map in Figure 2) are accumulated to them (illustrated in Appendix B). At certain threshold these cells will be filtered out as background by VOCA-NW while they can be correctly detected by VOCA.

We measured the same metrics as Sirinukunwattana et al. (2016) to quantitatively describe the accuracy of nuclei localization of VOCA. The Euclidean distance between each pair of ground truth and its assigned prediction was recorded for both folds of cross validation. The median, 1st quartile

and 3rd quartile of the distribution of the distances were reported. We emphasize again that the accurate coordinates of cell nuclei are the prerequisite for many downstream applications, such as tumor micro-environment analysis, and that low accuracy cell localization would result in accumulated error which hinders these tasks. Considering the radius of a cell nucleus is only around 6 to 12 pixels at  $20\times$  magnification, localization error of 5 pixels like [Sirinukunwattana et al. \(2016\)](#) may still introduce unignorable problems. VOCA explicitly learns nuclei localization via deep features and significantly reduced the error of 75% of the predictions to below 2.236 pixels.

## 6. Conclusion

In this paper, we proposed a novel deep learning algorithm called vector oriented confidence accumulation (VOCA) for large scale cell detection on histopathology images. The algorithm concurrently learns pixel-wise classification, localization and weight of contribution tasks that combine into an accumulator map which describes profoundly accurate and precise nuclei locations. Extensive experiments on a public cell detection dataset of colon cancer validated the efficacy of our proposed frame work and proved high detection performance and exceptional localization accuracy compared to the state-of-the-art, which implies high potential of a robust decision support application for various clinical and research purposes.

## Acknowledgements

This work was supported by the Warren Alpert Foundation Center for Digital and Computational Pathology at Memorial Sloan Kettering Cancer Center, the NIH/NCI Cancer Center Support Grant P30 CA008748, Weill Cornell Graduate School of Medical Sciences and the Tri-I Computational Biology and Medicine Program.

T.J.F. is the chief scientific officer, co-founder, and equity holder of Paige.AI.

## References

- Asif Ahmad, Amina Asif, Nasir Rajpoot, Muhammad Arif, et al. Correlation filters for detection of cellular nuclei in histopathology images. *Journal of medical systems*, 42(1):7, 2018.
- Yousef Al-Kofahi, Wiem Lassoued, William Lee, and Badrinath Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering*, 57(4):841–852, 2010.
- Sahirzeeshan Ali and Anant Madabhushi. An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery. *IEEE transactions on medical imaging*, 31(7):1448–1460, 2012.
- Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. Learning to detect cells using non-overlapping extremal regions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 348–356. Springer, 2012.
- Jianxu Chen and Chukka Srinivas. Automatic lymphocyte detection in h&e images with deep neural networks. *arXiv preprint arXiv:1612.03217*, 2016.

- Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer, 2013.
- Eric Cosatto, Matt Miller, Hans Peter Graf, and John S Meyer. Grading nuclear pleomorphism on histological micrographs. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.
- Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–410. Springer, 2013.
- Thomas J. Fuchs, Johannes Haybaeck, Peter J. Wild, Mathias Heikenwalder, Holger Moch, Adriano Aguzzi, and Joachim M. Buhmann. Randomized Tree Ensembles for Object Detection in Computational Pathology. In *Advances in Visual Computing: Part I, ISVC '09*, pages 367–378, Las Vegas, Nevada, 2009. ISBN 978-3-642-10330-8. doi: [http://dx.doi.org/10.1007/978-3-642-10331-5\\_35](http://dx.doi.org/10.1007/978-3-642-10331-5_35). URL [http://dx.doi.org/10.1007/978-3-642-10331-5\\_35](http://dx.doi.org/10.1007/978-3-642-10331-5_35).
- Ross Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015.
- Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.
- Jisoo Jeong, Hyojin Park, and Nojun Kwak. Enhancement of ssd by concatenating feature maps for object detection. *arXiv preprint arXiv:1705.09587*, 2017.
- Navid Alemi Koohababni, Mostafa Jahanifar, Ali Gooya, and Nasir Rajpoot. Nuclei detection using mixture density networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 241–248. Springer, 2018.
- H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. ISSN 1931-9193. doi: 10.1002/nav.3800020109. URL <http://dx.doi.org/10.1002/nav.3800020109>.
- Manohar Kuse, Yi-Fang Wang, Vinay Kalasannavar, Michael Khan, and Nasir Rajpoot. Local isotropic phase symmetry measure for detection of beta cells and lymphocytes. *Journal of pathology informatics*, 2, 2011.
- Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

- Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 850–855. IEEE, 2006.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop Autodiff*, 2017.
- Shan E Ahmed Raza, Khalid AbdulJabbar, Mariam Jamal-Hanjani, Selvaraju Veeriah, John Le Quesne, Charles Swanton, and Yinyin Yuan. Deconvolving convolution neural network for cell detection. *arXiv preprint arXiv:1806.06970*, 2018.
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- K. Sirinukunwattana, S. E. A. Raza, Y. W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, May 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2525803.
- Mohammad Tofighi, Tiantong Guo, Jairam KP Vanamala, and Vishal Monga. Deep networks with shape priors for nucleus detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 719–723. IEEE, 2018.
- Mitko Veta, Paul J Van Diest, Robert Kornegoor, André Huisman, Max A Viergever, and Josien PW Pluim. Automatic nuclei segmentation in h&e stained breast cancer histopathology images. *PloS one*, 8(7):e70221, 2013.
- Jelte Peter Vink, MB Van Leeuwen, CHM Van Deurzen, and G De Haan. Efficient nucleus detector in histopathology images. *Journal of microscopy*, 249(2):124–135, 2013.
- Haibo Wang, Angel Cruz Roa, Ajay N Basavanhally, Hannah L Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio Gonzalez, and Anant Madabhushi. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1(3):034003, 2014.
- W. Xie, J. A. Noble, and A. Zisserman. Microscopy cell counting with fully convolutional regression networks. In *MICCAI 1st Workshop on Deep Learning in Medical Image Analysis*, 2015a.

Yuanpu Xie, Fuyong Xing, Xiangfei Kong, Hai Su, and Lin Yang. Beyond classification: structured regression for robust cell detection using convolutional neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 358–365. Springer, 2015b.

Jun Xu, Lei Xiang, Qingshan Liu, Hannah Gilmore, Jianzhong Wu, Jinghai Tang, and Anant Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE transactions on medical imaging*, 35(1):119–130, 2016.

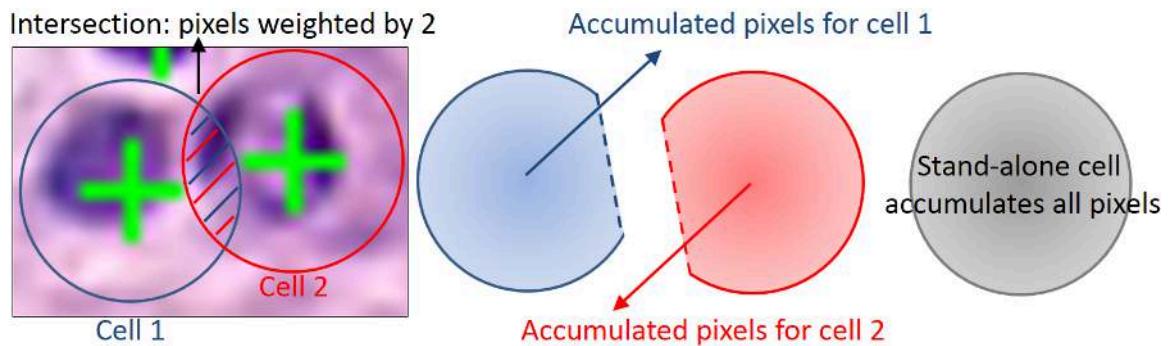
Yanning Zhou, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Sfcn-opi: Detection and fine-grained classification of nuclei using sibling fcn with objectness prior interaction. *arXiv preprint arXiv:1712.08297*, 2017.

## Appendix A. Image ids for each subsample

Subsample 1: 6, 8, 10, 11, 13, 17, 18, 19, 20, 21, 23, 25, 26, 27, 28, 29, 32, 33, 39, 41, 42, 45, 46, 47, 48, 49, 51, 53, 55, 56, 59, 60, 63, 65, 67, 69, 70, 75, 76, 78, 79, 84, 86, 87, 92, 93, 95, 96, 98, 100

Subsample 2: 1, 2, 3, 4, 5, 7, 9, 12, 14, 15, 16, 22, 24, 30, 31, 34, 35, 36, 37, 38, 40, 43, 44, 50, 52, 54, 57, 58, 61, 62, 64, 66, 68, 71, 72, 73, 74, 77, 80, 81, 82, 83, 85, 88, 89, 90, 91, 94, 97, 99

## Appendix B. Pixel sharing during confidence accumulation



# Unsupervised Lesion Detection via Image Restoration with a Normative Prior

**Suhang You**

JADENYOU1989@GMAIL.COM

**Kerem C. Tezcan**

TEZCAN@VISION.EE.ETHZ.CH

**Xiaoran Chen**

CHENX@VISION.EE.ETHZ.CH

**Ender Konukoglu**

ENDER.KONUKOGLU@VISION.EE.ETHZ.CH

*Computer Vision Laboratory, ETH Zurich, Sternwartstrasse 7, 8092 Zurich, Switzerland*

## Abstract

While human experts excel in and rely on identifying an abnormal structure when assessing a medical scan, without necessarily specifying the type, current unsupervised abnormality detection methods are far from being practical. Recently proposed deep-learning (DL) based methods were initial attempts at showing the capabilities of this approach. In this work, we propose an outlier detection method combining image restoration with unsupervised learning based on DL. A normal anatomy prior is learned by training a Gaussian Mixture Variational Auto-Encoder (GMVAE) on images from healthy individuals. This prior is then used in a Maximum-A-Posteriori (MAP) restoration model to detect outliers. Abnormal lesions, not represented in the prior, are removed from the images during restoration to satisfy the prior and the difference between original and restored images form the detection of the method. We evaluated the proposed method on Magnetic Resonance Images (MRI) of patients with brain tumors and compared against previous baselines. Experimental results indicate that the method is capable of detecting lesions in the brain and achieves improvement over the current state of the art.

**Keywords:** Unsupervised lesion detection, image restoration

## 1. Introduction

Identifying abnormal structures is an important component of radiological assessment. Arguably, abnormal structures, such as lesions, are first identified as outliers that do not fit expectations on normal anatomy, and then their types are specified. While the second task is tremendously difficult, even non-experts can excel in the first one. After showing a non-radiologist a small number of examples of images showing “normal” anatomy, they start to identify lesions with distinct intensity patterns as abnormal patterns.

Recently, research on supervised machine learning algorithms for lesion detection has taken huge strides in automated lesion detection of *prespecified* type, where models are optimized to detect lesions contained in a training set(Ayachi and Amor, 2009; Zikic et al., 2012; Geremia et al., 2011; Dong et al., 2017; Pereira et al., 2016; Kamnitsas et al., 2017; Li et al., 2018). Despite the success of supervised approaches, the problem of detecting any lesion, without specifying a type, as abnormal regions remains a very challenging problem. Advancing on this task would facilitate new applications in acquisition and screening. Furthermore, unlike supervised methods, unsupervised methods may not require extensive datasets from patients, making them attractive from a practical point of view.

Over the last two decades, many unsupervised lesion detection methods have been proposed. Prior to DL-based models, Van Leemput *et al* (Van Leemput et al., 2001) utilized registration to a healthy brain atlas and mixture models based on tissue-specific intensity to detect lesions, followed by Moon *et al* (Moon et al., 2002) using an atlas with spatial features instead and Prastawa *et al* (Prastawa et al., 2004) combining spatial and intensity atlases. More recent non-DL works moved from modeling pixels independently to modeling small image patches through dimensionality reduction methods such as principle component analysis (Zacharaki and Bezerianos, 2012) and (Erus et al., 2014), statistical patch-wise representations in (Cardoso et al., 2015) and sparse representation (Zeng et al., 2016). In the latest years, neural network based generative models such as Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), and Variational Auto-encoders (Kingma and Welling, 2013) (VAE) have been applied to unsupervised lesion detection. The underlying approach, similar to non-DL methods, is to first learn a normative distribution using GAN or AE-based methods and then detect areas with low-likelihood as lesions, with respect to the distribution. The common detection procedure is to project a given image to the latent space of the learned model, reconstruct it and identify differences between reconstruction and the original image. Areas not fitting to the normative distribution cannot be reconstructed faithfully, leading to high reconstruction errors. AnoGAN (Schlegl et al., 2017) used a GAN within the described common approach. The projection to the latent space was solved as an optimization problem, seeking the best latent space representation for a given image. The optimization can be difficult to solve in practice. Methods proposed in (Baur et al., 2018; Pawlowski et al., 2018; Chen and Konukoglu, 2018) used AE-based methods to learn the normative distribution, which facilitated projection to the latent space.

In this work, we propose a novel unsupervised lesion detection method based on image restoration. The goal of the method is to identify a pixel-wise detection map like a segmentation. Similar to previous approaches, the framework consists of approximating a normative prior distribution of healthy images using an AE-based method, specifically Gaussian Mixture Variational Autoencoder (GMVAE) (Dilokthanakul et al., 2016; Johnson et al., 2016). Different than previous methods, we cast the detection as an image restoration problem and solve it by Maximum-A-Posteriori (MAP) estimation using the normative distribution as the prior term, similar to the method proposed in (Tezcan et al., 2018) for image reconstruction. The detection method is iterative and leverages the differentiability of the normative distribution modeled via neural networks. The final detection result is given as the difference between restored and original image. We evaluate the proposed method using MRI from patients with brain tumors and compare with other methods.

## 2. Methods

The proposed method consists of two steps : 1) modelling the distribution of healthy images, i.e. the normative distribution, using a GMVAE and 2) detecting lesions as outliers with MAP restoration. First we briefly explain GMVAE, then continue with the restoration algorithm and finally give some details on parameter selection.

### 2.1. GMVAE

We first introduce the VAE model and then GMVAE as its extension.

VAE (Kingma and Welling, 2013) is an unsupervised density estimation method that approximates the distribution of high dimensional data, e.g. images, from a given set of samples. It for-

mulates a latent variable model, written as  $\log p(X) = \log \int p(X|z)p(z)dz$ , with  $X \in \mathbb{R}^N$  the image,  $z \in \mathbb{R}^L$  the latent variable ( $N \gg L$ ) and  $p(z)$  a pre-specified prior on the latent variable, where  $p(X|z)$  is modeled as a neural network. The learning is cast as maximizing the log-likelihood,  $\log p(X)$ , of the observed samples. The direct evaluation of the integral is often not analytically achievable. As a remedy, a proposal distribution  $q(z|X)$  that approximates the true posterior  $p(z|X)$  is introduced. This allows formulating a lower bound to  $\log p(X)$  as the evidence lower bound (ELBO):

$$\log p(X) \geq \text{ELBO} = \mathbb{E}_{q(z|X)}[\log p(X|z)] - KL[q(z|X)||p(z)], \quad (1)$$

where  $KL$  denotes the Kullback-Leibler divergence. The first term is a reconstruction loss,  $X$  is first projected to the latent space and then back-projected to the image space, i.e. reconstructed, and  $\mathbb{E}_{q(z|X)}[\log p(X|z)]$  measures the expected deviation of the reconstruction and observation. The second term measures the divergence between the encoded distribution and the prescribed prior, acting as a regularizer. The equation becomes an equality when  $q(z|X) = p(z|X)$ . In VAEs,  $q(z|X)$ , the encoder, can be modeled as a Gaussian  $N(\mu_z(X), \sigma_z^2(X)\mathbf{I})$ , where  $\mu_z(X)$  and  $\sigma_z(X)$  are functions of the input  $X$  that are parameterized by a neural network. Similarly,  $p(X|z)$ , the decoder, can also be modeled as a Gaussian  $N(\mu_X(z), \sigma_X^2(z)\mathbf{I})$  where variables are functions of  $z$  parameterized by another network. For both Gaussians, diagonal covariance matrices are used,  $\mathbf{I}$  represent the identity matrix with appropriate dimensions. The training then aims at optimizing network parameters to maximize the ELBO, for a given set of training samples. After training, the ELBO becomes a close approximation to the true distribution  $p(X)$ .

**GMVAE** (Dilokthanakul et al., 2016) replaces the unit Gaussian prior on the latent space with a Gaussian mixture model, which leads to higher representation power. GMVAE has three latent variables  $z, \omega, c$  and models  $p(z|\omega, c) = \prod_{k=1}^K N(\mu_{c_k}(\omega), \text{diag}(\sigma_{c_k}^2(\omega)))^{c_k}$ . Here,  $K$  is the pre-specified number of components,  $c \sim \text{Mult}(\frac{1}{K})$  is a one-hot vector and  $\omega \sim N(0, I)$ .  $z|\omega$  is a Gaussian mixture model and the parameters  $\mu_{c_k}(\omega), \sigma_{c_k}^2(\omega)$  are functions of  $\omega$  parameterized as neural networks.

In the GMVAE formulation, the ELBO is expressed as

$$\begin{aligned} & \mathbb{E}_{q(z|X)}[\log p(X|z)] - \mathbb{E}_{q(\omega|X)p(c|z,\omega)}[KL[q(z|X)||p(z|\omega, c)]] \\ & - KL[q(\omega|X)||p(\omega)] - \mathbb{E}_{q(z|X)q(\omega|X)}[KL[p(c|z, \omega)||p(c)]], \end{aligned} \quad (2)$$

where the first term is the same reconstruction term as in Eqn. 1. When the ELBO is maximized, the second term ensures that the encoder distribution fits the prior, the third term makes sure the posterior of  $\omega$  does not diverge much from the prior  $p(\omega)$  and the last term enforces that the model does not collapse into a single Gaussian but uses the mixture model. Similar to VAE, all the probability functions are parameterized by networks and GMVAE model is trained to maximize the ELBO for a set of training samples. As its latent space has higher modeling capacity, the GMVAE should in theory make a better approximation to  $p(X)$  than the vanilla VAE after the training. We use 7 convolutional layers for the encoder and 7 transpose convolutional layers for the decoder. The latent variables  $z$  and  $\omega$  are implemented as 2D structures with sizes 32x42x1. We use  $c = 9$  clusters. For further details on the architectures we refer the reader to the Appendix A.

## 2.2. Restoration of the image with lesions

Here, we assume that a normative distribution  $p(X)$  is learned with an AE-based model using images from healthy individuals. Although our final method uses GMVAE, the proposed detection process

is generic and can be applied to other AE-based models, hence, we present it considering its generality. Denoting an image with lesion with  $Y \in \mathbb{R}^N$ , we model the lesion as an additive component, and  $Y$  can then be written as  $Y = X + \hat{D}$ , with  $\hat{D}$  being the lesion and  $X$  the ‘healthy’ counterpart of  $Y$  without the lesion.  $\hat{D}$  here is a pixel-wise lesion image, which the proposed algorithm aims to determine.

The goal of the restoration is to find  $X$  given  $Y$ , i.e. to find the corresponding healthy image where the healthy region in  $Y$  is unchanged and lesion region is replaced by ‘healthy’ structures. We model the restoration as the following MAP estimation problem

$$\arg \max_X \log p(X|Y) = \arg \max_X [\log p(Y|X) + \log p(X)] \quad (3)$$

The first term  $p(Y|X)$  stands for a data consistency term, which we detail in Section 2.2, and the second term is the normative prior learned from healthy images. In our proposed method, we use the ELBO from the trained GMVAE to approximate the prior distribution as was done in (Tezcan et al., 2018). In this case, the equation can be reformulated as

$$\arg \max_X \log p(X|Y) \approx \arg \max_X [\log p(Y|X) + ELBO(X)]. \quad (4)$$

Since the prior model is trained on healthy images, it will assign high probabilities to such images and low probability to images with lesions. When modified to maximize the ELBO, an image with lesions appears more similar to a healthy image with lesions removed during the process. However, in Eqn 4, while the ELBO tries to change the image, the data consistency term prevents big changes, resulting in a balanced optimization problem.

We restore the anomalous images using gradient ascent. Taking the derivative of Eqn (4) w.r.t. to input  $X$  we obtain  $G(x) = \frac{d}{dx} [\log p(Y|X) + ELBO(X)]|_{X=x}$ , where the ELBO term is also differentiable with respect to its input (Tezcan et al., 2018). The iterative gradient ascent equation is given as usual as  $X_{i+1} = X_i + \alpha_i \cdot G(X_i)$ , where  $i$  is the iteration index and  $\alpha_i$  is the step size. We initialize the input images as  $X_0 = Y$  and after  $n$  steps we have the restored image  $\hat{X} = X_n$ . The pixel-wise lesion map is then given as  $\hat{D} = Y - \hat{X}$ . Considering that lesion effects can be both negative and positive, the final detection of the proposed method is given as  $D = |\hat{D}| = |Y - \hat{X}|$ .

During training, the prior model learns to assign high standard deviation  $\sigma_z(X)$  to regions where the reconstruction mean  $\mu_z(X)$  deviates from the samples, and lesions correspond to precisely such regions. Since the gradient incorporates the inverse of standard deviation as a factor, the magnitude in lesions get down-scaled, causing instability issues. We avoid this by setting the standard deviation of  $p(X|z)$  to  $1/\sqrt{2}$  in the whole image during restoration. This modification is heuristic and not ideal, but it empirically works on the hold-out validation set. A summary of the restoration process is presented in Appendix D.

**Data Consistency** The likelihood  $p(Y|X)$  measures the distance between  $X$  and  $Y$  and serves as the data consistency term, which punishes deviations in  $X$  from  $Y$ .

We assume that lesions are structurally compact areas that can be modeled with piece-wise linear functions with sparse gradients, however, we cannot tell anything about their intensity values. To incorporate this assumption into our model, we use the TV norm (Rudin et al., 1992), where the lesion  $\hat{D} = Y - X$  is assumed to have a low TV norm and  $\hat{X} = \arg \max_X [-\lambda ||X - Y||_{TV} + ELBO(X)]$ , where  $\lambda > 0$ , is a multiplier and balances the effect of the TV norm and the ELBO. Unfortunately, it is not trivial to set  $\lambda$  in the unsupervised scenario.

Below, we present a heuristic method for determining  $\lambda$  based on the training data.

### 2.3. Determining $\lambda$ for the data consistency term

In the unsupervised setting we have no access to images with lesions, therefore, we cannot rely on any such images to determine  $\lambda$ . Instead, we choose the  $\lambda$  value based on the changes it causes on the training images. As all the training images are acquired from healthy individuals with no lesions, ideally the restoration framework should yield no change on them. However, due to the approximate nature of the normative distribution modeled with networks and possible issues with optimization, slight changes will occur. For a wide range of  $\lambda$  values, we compute the average change incurred on the training set and use the  $\ell_1$  distance to quantify the average change and calculate  $\varepsilon(\lambda) = \frac{1}{M} \sum_{\{Y_s\}} |Y_s - \hat{X}_s^\lambda|$ , which is consistent with the metric used to detect outliers.

Here  $M$  is defined as the number of subjects in  $\{Y_s\}$ . In Figure 2b, we plot  $\varepsilon(\lambda)$  vs  $\lambda$  computed over a set of 52 healthy subjects, for which details are given in Section 3. We observe that  $\varepsilon(\lambda)$  has a minimum value at a certain  $\lambda$ . We choose the  $\lambda$  value that yields the minimum change in the set of training images. Note that the curve does not decrease as  $\lambda$  increases because (i)  $\lambda$  weights the TV norm, which is different from  $\varepsilon(\lambda)$ , and (ii) optimization is possibly non-convex and certain  $\lambda$  might yield more faithful restorations, which is also important for lesion detection.

### 2.4. Calculating the threshold for masking

With the restored image  $\hat{X}$  and the input image  $Y$ , the residual image  $D$  is calculated as  $D = |\hat{X} - Y|$ . To create a binary detection map  $S$  from  $D$ , we need to identify a threshold  $T_{ls}$ , where pixels with  $D$  values larger than  $T_{ls}$  will be labeled as lesion and others as healthy.

As we do not have ground truth detections in the unsupervised scenario, we use the approach proposed in (Konukoglu et al., 2018) for determining the threshold.

We assume all training images are lesion free and therefore any detection with the proposed method are false positives. Based on this, we set a limit for the False Positive Rate (FPR)  $l_{FPR}$  in the training set and determine the minimum threshold on  $D$  that satisfies the limit.

We obtain the threshold by the golden section search algorithm (Kiefer, 1953) to convert the residual maps  $D$  to binary detection maps  $S$  and calculate the Dice coefficient (DSC) for each test subject.

To provide a baseline for the DSC values, we calculate DSC\_AUC using a threshold obtained by ROC curves. Specifically, We use ROC curves to select the threshold that leads to the maximum value of  $TPR - FPR$ , and calculate DSC for each subject based on this threshold. This yields an optimistic score and not used to assess the model.

## 3. Experimental Settings

### 3.1. Datasets & Preprocessing

The current implementation of our method applies to 2D slices. Specifically, we used T2-weighted MR images from the Cambridge Centre for Ageing and Neuroscience dataset (CamCANT2) (Taylor et al., 2017) to train our network. It contains 652 subjects with lesion-free brain slices where 600 subjects were randomly selected as the training data and 52 as test data. We performed lesion detection on T2 weighted images from the Multimodal Brain Tumor Image Segmentation (BRATS) (Menze et al., 2015) Challenge 2017 dataset.

For both datasets, we first normalized image intensities by computing  $\frac{I - \min(I)}{\max(I) - \min(I)}$  for each subject, where  $I$  are pixel intensities. We trained and tested our method on transversal slices. Since

CamCANT2 and BRATS datasets have different intensity characteristics, we performed histogram matching on BRATS where we randomly chose a subject from CamCANT2 dataset and matched the histogram of each BRATS subject to it. Lastly, we excluded slices with only background and cropped excessive background of all slices for both training and lesion detection to a size of  $158 \times 198$  to accelerate computation.

### 3.2. Implementation Details

**General Settings:** We trained a GMVAE and ran the restoration for 500 steps in all experiments. Details for network architectures and model training are presented in Appendix B. We experimentally explored the effect of different number of restoration steps and the number of clusters. We used the optimal cluster number for detection  $c = 9$ .

**Selecting the optimal  $\lambda$  and threshold values:** We used the 52 subjects from the CamCANT2 test dataset to determine  $\lambda$  value and the threshold as in section 2.3 and 2.4.

The  $\lambda$  values and the thresholds were then used for the evaluation. In order to understand the sensitivity of the proposed method to different  $\lambda$  values, we additionally ran restoration with different values and present the results.

## 4. Results

### 4.1. Overall lesion detection result

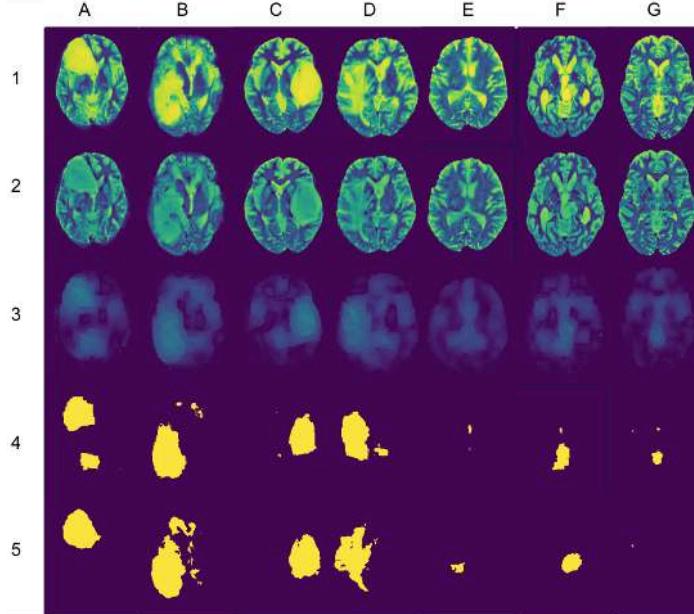


Figure 1: Segmentation by GMVAE(TV) at DSC5. Top to bottom: images with lesions, restored images, residual images, predicted segmentations, groundtruth segmentations.

Visual results in Figure 1 show seven randomly selected detection results at DSC5. Illustration for DSC1, DSC10 and DSC\_AUC are provided in Appendix E. The proposed method is able to detect lesions, especially when the lesions are of a relatively large size or appear in high intensity such as in columns A to D. On the other hand, the method has difficulty detecting smaller lesions as shown in columns E to G.

Table 1 presents the evaluated metrics for baselines and our method. GMVAE performs the best in terms of DSC\_AUC, AUC, DSC5 and DSC10 among all the methods. Compared to AnoGAN (Schlegl et al., 2017), VAE-256, VAE-128 and AAE-128 (Chen et al., 2018), we achieve respectively 28%, 24%, 20% and 19% increase in AUC.

Table 1: Summarized AUC and DSC for GMVAE(TV) and baseline methods. FPR and FNR are calculated from  $T_{ls}$  at DSC\_AUC. DSC1, DSC5, DSC10 are calculated from  $T_{lss}$  at  $l_{FPR} = 0.01, 0.05, 0.10$ . For GMVAE(TV),  $\lambda = 1.8$ . na: not available.

Methods	DSC_AUC	AUC	FPR	FNR	DSC1	DSC5	DSC10
VAE(TV) (ours)	0.34±0.18	0.80	<b>0.11</b>	0.40	<b>0.34±0.20</b>	0.36±0.27	0.40±0.24
GMVAE(TV) (ours)	<b>0.37±0.18</b>	<b>0.83</b>	0.12	<b>0.34</b>	0.22±0.21	<b>0.46±0.23</b>	<b>0.43±0.20</b>
VAE-256	na	0.67	0.26	0.43	na	na	na
VAE-128	0.22±0.14	0.69	0.21	0.46	0.09±0.06	0.19±0.15	0.26±0.17
AAE-128	0.23±0.13	0.70	0.25	0.43	0.03±0.03	0.18±0.14	0.23±0.15
AnoGAN	0.19±0.10	0.65	0.33	0.37	0.02±0.02	0.10±0.06	0.19±0.13

We also plot the Receiver operating characteristic (ROC) curves in Figure 2c using the whole dataset. The ROC curves are consistent with the AUC values shown in Table 1.

#### 4.2. Sensitivity to $\lambda$ values

In Figure 2a, we show the AUC values for the proposed GMVAE(TV) within the range of [0.6, 4.0]. The performance of the method appears to be relatively stable, which is desirable due to the difficulty of parameter selection in the unsupervised setting.

In Figure 2b we show the change of  $\varepsilon(\lambda)$  using different  $\lambda$ s for GMVAE(TV) restoration. The lowest  $\varepsilon(\lambda)$  value is obtained with  $\lambda = 1.8$ .

#### 4.3. Sensitivity to number of clusters and restoration steps:

We present results for varying number of clusters, shown in Table 2. The results suggest that, although the performance may change, the method works for all number of clusters we experimented with. We also show results for AUC values vs number of restoration steps used in Appendix C for the whole dataset, which indicate convergence at 500 steps.

### 5. Conclusion

In this paper, we proposed an unsupervised lesion detection method based on image restoration with a normative prior learned via AE-based neural network models, specifically GMVAE. The result showed that our method was able to detect brain tumors without any supervision, achieving

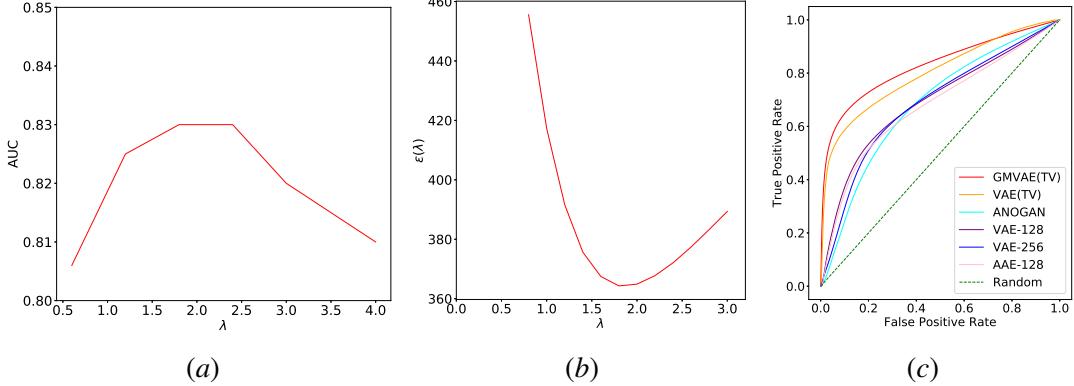


Figure 2: (2a) AUC vs.  $\lambda$ . The AUC values are calculated using all subjects in BRATS dataset. (2b)  $\varepsilon(\lambda)$  vs.  $\lambda$ . (2c) ROC curves on BRATS dataset.

Table 2: AUC/DSC values for varying number of clusters. Mean  $\pm$  std are shown for DSC. FPR and FNR are calculated from  $T_{ls}$  at DSC\_AUC. DSC1, DSC5, DSC10 are calculated from  $T_{lss}$  at  $l_{FPR} = 0.01, 0.05, 0.1$

Cluster Size	DSC_AUC	AUC	FPR	FNR	DSC1	DSC5	DSC10
c = 3	0.27 $\pm$ 0.16	0.78	0.20	0.35	0.04 $\pm$ 0.07	0.26 $\pm$ 0.19	0.32 $\pm$ 0.19
c = 6	0.30 $\pm$ 0.18	0.73	<b>0.11</b>	0.49	0.06 $\pm$ 0.13	0.35 $\pm$ 0.21	0.28 $\pm$ 0.17
<b>c = 9</b>	<b>0.37<math>\pm</math>0.18</b>	<b>0.83</b>	0.12	<b>0.34</b>	<b>0.22<math>\pm</math>0.21</b>	<b>0.46<math>\pm</math>0.23</b>	<b>0.43<math>\pm</math>0.20</b>
c = 12	0.30 $\pm$ 0.17	0.77	0.14	0.43	0.08 $\pm$ 0.13	0.37 $\pm$ 0.22	0.33 $\pm$ 0.18

high AUCs and DSCs, improving on the state-of-the-art methods. Further experiments revealed that the model is robust to parameter selection to a reasonable extent. This article presents the technical details and further research on applying the methodology on different lesions will prove its value as a general purpose unsupervised lesion detection method.

## Acknowledgments

We thank Swiss National Science Foundation for financially supporting this work. We also thank NVIDIA for their generous GPU donations.

## References

- Raouia Ayachi and Nahla Ben Amor. Brain tumor segmentation using support vector machines. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 736–747. Springer, 2009.

- Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. *arXiv preprint arXiv:1804.04488*, 2018.
- M Jorge Cardoso, Carole H Sudre, Marc Modat, and Sebastien Ourselin. Template-based multi-modal joint generative model of brain data. In *International Conference on Information Processing in Medical Imaging*, pages 17–29. Springer, 2015.
- Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018.
- Xiaoran Chen, Nick Pawlowski, Martin Rajchl, Ben Glocker, and Ender Konukoglu. Deep generative models in the real-world: An open challenge from medical imaging. *arXiv preprint arXiv:1806.05452*, 2018.
- Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.
- Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo, and Yike Guo. Automatic brain tumor detection and segmentation using u-net based fully convolutional networks. In *Annual Conference on Medical Image Understanding and Analysis*, pages 506–517. Springer, 2017.
- Guray Erus, Evangelia I Zacharakis, and Christos Davatzikos. Individualized statistical learning from medical image databases: Application to identification of brain lesions. *Medical image analysis*, 18(3):542–554, 2014.
- Ezequiel Geremia, Olivier Clatz, Bjoern H Menze, Ender Konukoglu, Antonio Criminisi, and Nicholas Ayache. Spatial decision forests for ms lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, 57(2):378–390, 2011.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- Matthew J Johnson, David Duvenaud, Alexander B Wiltschko, Sandeep R Datta, and Ryan P Adams. Structured vaes: Composing probabilistic graphical models and variational autoencoders. *arXiv preprint arXiv:1603.06277*, 2, 2016.
- Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.
- Jack Kiefer. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3):502–506, 1953.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ender Konukoglu, Ben Glocker, Alzheimer's Disease Neuroimaging Initiative, et al. Reconstructing subject-specific effect maps. *NeuroImage*, 181:521–538, 2018.

Hongwei Li, Gongfa Jiang, Jianguo Zhang, Ruixuan Wang, Zhaolei Wang, Wei-Shi Zheng, and Bjoern Menze. Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images. *NeuroImage*, 183:650–665, 2018.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993, 2015.

Nathan Moon, Elizabeth Bullitt, Koen Van Leemput, and Guido Gerig. Automatic brain and tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 372–379. Springer, 2002.

Nick Pawlowski, Matthew C. H. Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan K. Stevenson, Aneesh M Khetani, Tom Newman, Fred A Zeiler, Richard John Digby, Jonathan P Coles, Daniel Rueckert, David K. Menon, Virginia F. J. Newcombe, and Ben Glocker. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders. In *International Conference on Medical Imaging with Deep Learning (MIDL 2018) - Abstracts Track*, 2018.

Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.

Marcel Prastawa, Elizabeth Bullitt, Sean Ho, and Guido Gerig. A brain tumor segmentation framework based on outlier detection. *Medical image analysis*, 8(3):275–283, 2004.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.

Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The cambridge centre for ageing and neuroscience (cam-can) data repository: structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage*, 144:262–269, 2017.

Kerem C Tezcan, Christian F Baumgartner, Roger Luechinger, Klaas P Pruessmann, and Ender Konukoglu. Mr image reconstruction using deep density priors. *IEEE transactions on medical imaging*, 2018.

Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, Alan Colchester, and Paul Suetens. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE transactions on medical imaging*, 20(8):677–688, 2001.

Evangelia I Zacharaki and Anastasios Bezerianos. Abnormality segmentation in brain images via distributed estimation. *IEEE Transactions on Information Technology in Biomedicine*, 16(3):330–338, 2012.

Ke Zeng, Guray Erus, Aristeidis Sotiras, Russell T Shinohara, and Christos Davatzikos. Abnormality detection via iterative deformable registration and basis-pursuit decomposition. *IEEE transactions on medical imaging*, 35(8):1937–1951, 2016.

D Zikic, B Glocker, E Konukoglu, J Shotton, A Criminisi, D Ye, C Demiralp, OM Thomas, T Das, R Jena, et al. Context-sensitive classification forests for segmentation of brain tumor tissues. *Proc MICCAI-BraTS*, pages 1–9, 2012.

## Appendix A. Network Architecture and Training Details

We build the GMVAE as a fully convolutional network and the latent space as a 2D structure. The network is shown in Table 3. The encoder consist of 7 convolutional layers and the decoder consist of 5 transposed convolutional layers and 2 convolutional layers.

We set the dimension of latent variables  $z = 32 \times 42 \times 1$ , the dimension of the prior  $\omega = 32 \times 42 \times 1$  and the number of mixtures  $c = 9$ .

Table 3: The encoder encodes  $q(z|X)$  and  $q(\omega|X)$ . They share layers except the output where four  $1 \times 1$  convolution layers are used. The layers are connected from top to bottom.  $\{*\}_n$  means the layer with the given settings are used  $n$  times. Conv(\*) is the convolutional layer.  $\mu_*$  and  $\sigma_*$  are means and standard deviations of output variables  $z$ ,  $\omega$  and  $X$ . ReLU means rectified linear unit activation function. The decoder decodes  $p(X|z)$ . Upconv(\*) is the transposed convolutional layer. tf.image.resize\_images is a built in function of Tensorflow for resizing images, where we use nearest neighbor method. Network  $p(z|\omega, c)$  is to generate distributions of  $z$  given the prior  $\omega$  and mixture category  $c$

Structure	Input	Layers	Output
Encoder $q(z X)$ and $q(\omega X)$	$X$ $158 \times 198 \times 1$	$\{\text{Conv}(3 \times 3 \times 64, \text{stride} = 2), \text{ReLU}\}_1$ $\{\text{Conv}(3 \times 3 \times 64, \text{stride} = 1), \text{ReLU}\}_2$ $\{\text{Conv}(3 \times 3 \times 64, \text{stride} = 2), \text{ReLU}\}_1$ $\{\text{Conv}(3 \times 3 \times 64, \text{stride} = 1), \text{ReLU}\}_2$ $\{\text{Conv}(1 \times 1 \times 1, \text{stride} = 1)\}_1$	$\mu_z$ and $\sigma_z$ $32 \times 42 \times 1$ $\mu_\omega$ and $\sigma_\omega$ $32 \times 42 \times 1$
Decoder $p(X z)$	$z$ $32 \times 42 \times 1$	$\{\text{UpConv}(1 \times 1 \times 64, \text{stride} = 1), \text{ReLU}\}_1$ $\{\text{UpConv}(3 \times 3 \times 64, \text{stride} = 1), \text{ReLU}\}_2$ tf.image.resize_images, Upsampling $\times 2$ $\{\text{Conv}(3 \times 3 \times 64, \text{stride} = 1), \text{ReLU}\}_1$ $\{\text{UpConv}(3 \times 3 \times 64, \text{stride} = 1), \text{ReLU}\}_2$ tf.image.resize_images, Upsampling $\times 2$ $\{\text{Conv}(3 \times 3 \times 1, \text{stride} = 1)\}_1$	$\mu_X$ $158 \times 198 \times 1$ $\sigma_X$ $158 \times 198 \times 1$
Network $p(z \omega, c)$	$\omega$ $32 \times 42 \times 1$	$\{\text{Conv}(1 \times 1 \times 64, \text{stride} = 1), \text{ReLU}\}_1$ $\{\text{Conv}(1 \times 1 \times 1, \text{stride} = 1)\}_1$	$\mu_z \omega, c$ and $\sigma_z \omega, c$ $32 \times 42 \times 1$

## Appendix B. Training and Restoration details

To train the VAE and GMVAE, we use the Adam optimizer with parameters  $\beta_1=0.9$ ,  $\beta_2=0.999$ ,  $\epsilon = 1 \times 10^{-8}$ , and a learning rate of  $5 \times 10^{-5}$ . For restoration, in both evaluation and while determining the  $\lambda$  values, we run the gradient ascent for  $n = 500$  iterations with a learning rate of  $\alpha = 1 \times 10^{-3}$ .

We use Tensorflow to implement the network as well as the restoration procedure. The implementation code can be found at <https://github.com/yousuhang/Unsupervised-Lesion-Detection-via-Image-Restoration-with-a-Normative-Prior>

### Appendix C. AUC values vs. restoration step

Here we show the convergence of performance in terms of AUC for the whole dataset with increasing number of restoration steps.

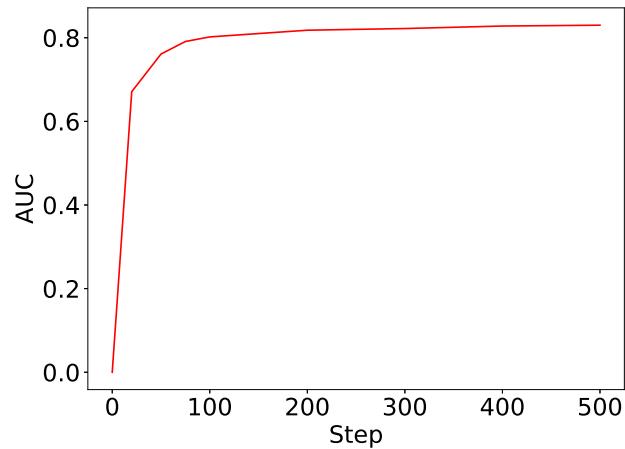


Figure 3: AUC vs Restoration Step plot

## Appendix D. Lesion Detection Algorithm

Here we present a summary of the proposed algorithm for lesion detection.

---

### Algorithm 1: Unsupervised Lesion Detection

**Input:**  $Y'$ : original image with lesion;  $ELBO$ : a GMVAE trained only on healthy images;  $X$ : a representative image from the training set for histogram matching ;  $T_{ls}$ : threshold for masking;  $\alpha_i$ : step size for each iteration

**Output:**  $S$ : Predicted Detection

**Procedure** DETECT( $Y'$ ,  $ELBO$ ,  $X$ ,  $T_{ls}$ ,  $\alpha_i$ )

```

 $Y \leftarrow HistEq(Y', X)$            // fit histogram of  $Y'$  to histogram of  $X$ .
 $X_0 \leftarrow Y$                    // initialize  $X_i$  with the equalized test image.
for  $i = 0$  to  $n - 1$  do           // restoration iterations

     $G(X_i) \leftarrow \frac{d}{dX} [\log p(Y|X) + ELBO(X)]|_{X=X_i}$ 
     $X_{i+1} \leftarrow X_i + \alpha_i \cdot G(X_i)$            // update image using the gradient
end
 $\hat{X} \leftarrow X_n$                  // restored image
 $D \leftarrow |Y - \hat{X}|$            // calculate the residual image
 $S \leftarrow \text{threshold}(D, T_{ls})$      // threshold residual images to obtain lesion
                                         labels
return  $S$                       // Resulting segmentation map

```

---

## Appendix E. Other Visual Results

Here we show other segmentation examples for thresholds chosen at DSC1 (Figure 4), DSC10 (Figure 5) and DSC\_AUC (Figure 6). The input images are the same as in Figure 1.

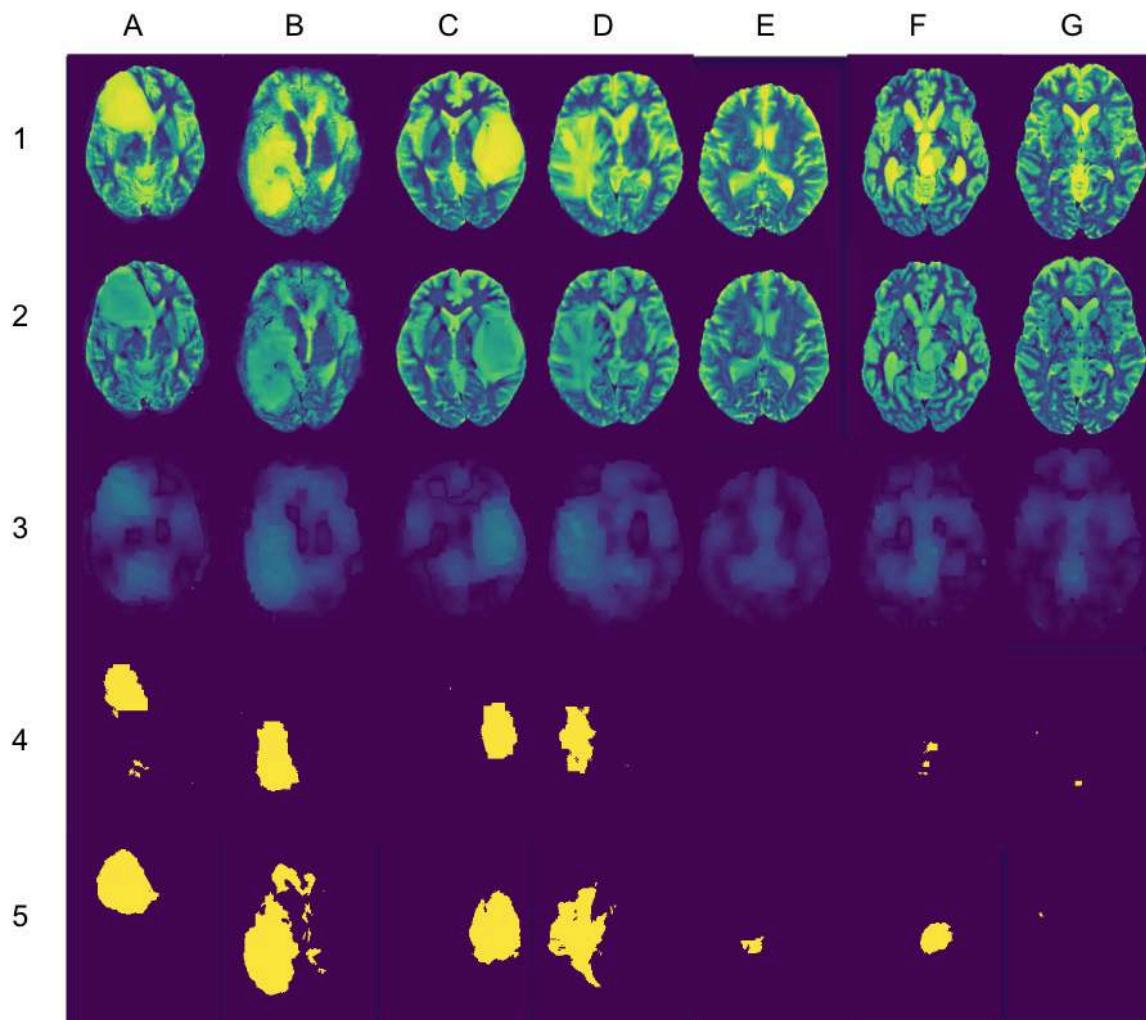


Figure 4: BRATS segmentation results by GMVAE(TV) for DSC1. Row 1: images with lesions; row 2~4: restored images, residual images and segmentations; row 5: ground truth segmentations.

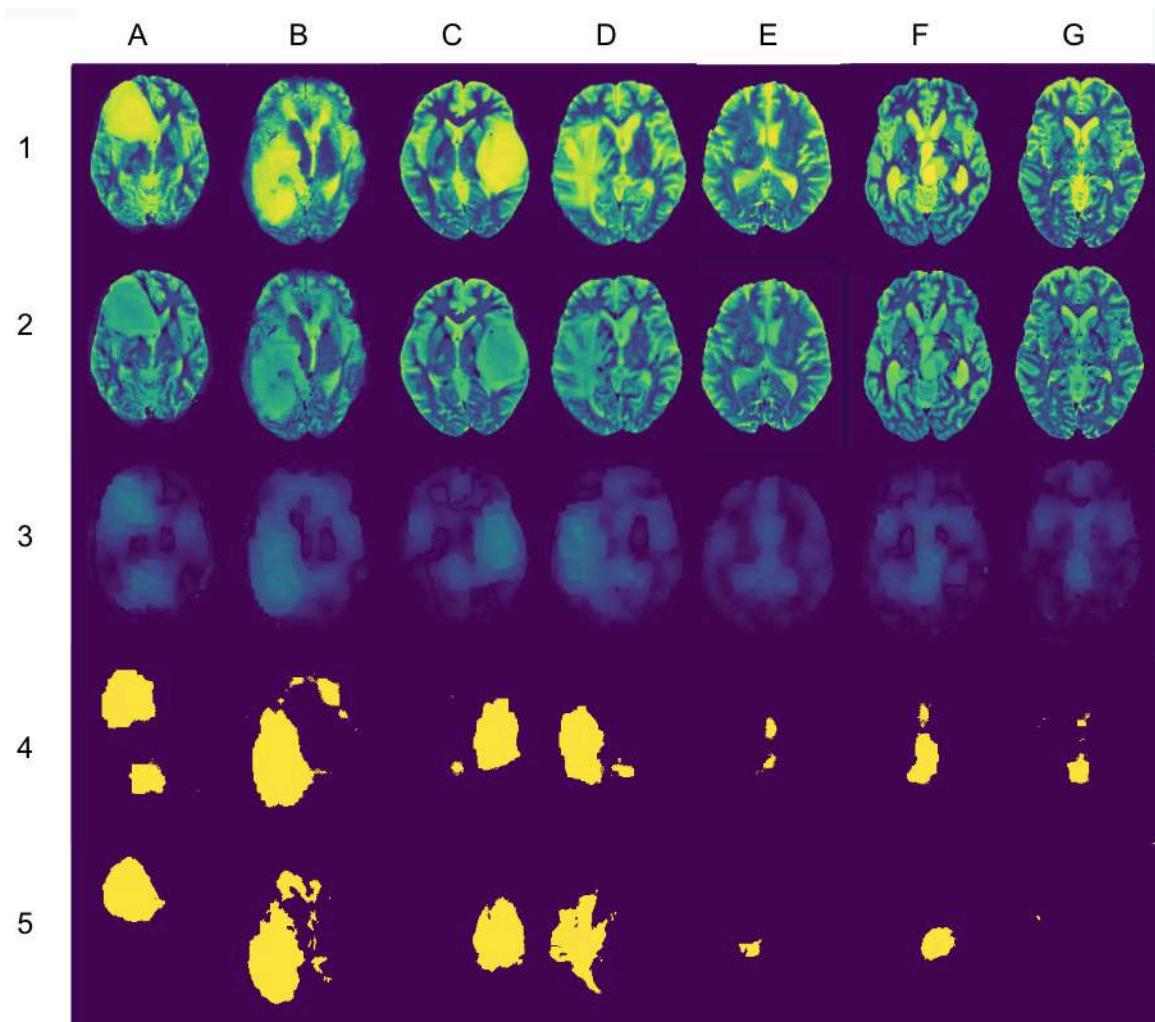


Figure 5: BRATS segmentation results by GMVAE(TV) for DSC10. Row 1: images with lesions; row 2~4: restored images, residual images and segmentations; row 5: ground truth segmentations.

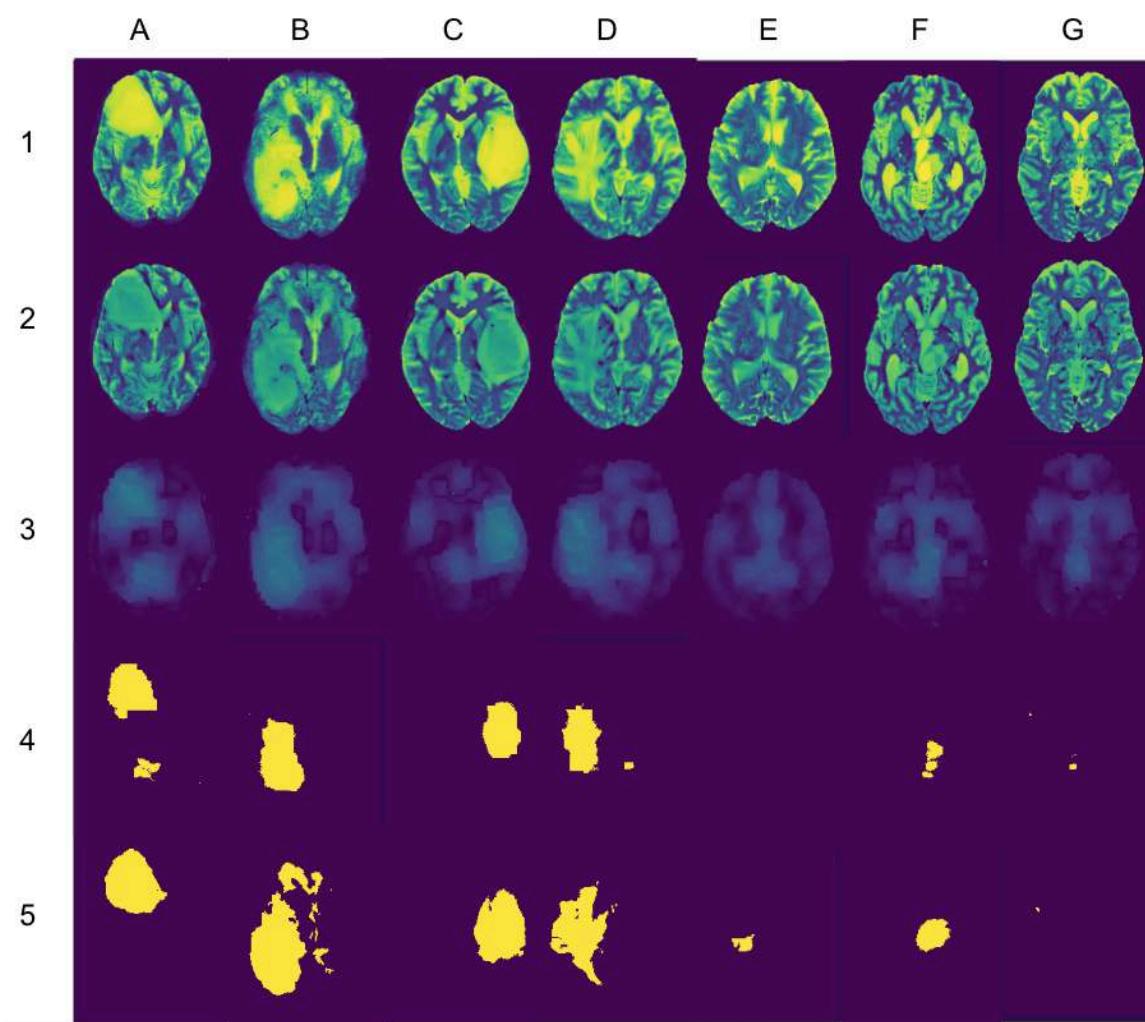


Figure 6: BRATS segmentation results by GMVAE(TV) for DSC\_AUC. Row 1: images with lesions; row 2~4: restored images, residual images and segmentations; row 5: ground truth segmentations.

# Deep Learning Approach to Semantic Segmentation in 3D Point Cloud Intra-oral Scans of Teeth

**Farhad Ghazvinian Zanjani<sup>1</sup>**

F.GHAZVINIAN.ZANJANI@TUE.NL

**David Anssari Moin<sup>2</sup>**

DAVID@PROMATON.COM

**Bas Verheij<sup>2</sup>**

BAS@PROMATON.COM

**Frank Claessen<sup>2</sup>**

FRANK@PROMATON.COM

**Teo Cherici<sup>2</sup>**

TEO@PROMATON.COM

**Tao Tan<sup>1</sup>**

T.TAN1@TUE.NL

**Peter H. N. de With<sup>1</sup>**

P.H.N.DE.WITH@TUE.NL

<sup>1</sup> *Eindhoven University of Technology, 5600MB Eindhoven, The Netherlands*

<sup>2</sup> *Promaton Inc., 1076GR Amsterdam, The Netherlands*

## Abstract

Accurate segmentation of data, derived from intra-oral scans (IOS), is a crucial step in a computer-aided design (CAD) system for many clinical tasks, such as implantology and orthodontics in modern dentistry. In order to reach the highest possible quality, a segmentation model may process a point cloud derived from an IOS in its highest available spatial resolution, especially for performing a valid analysis in finely detailed regions such as the curvatures in border lines between two teeth. In this paper, we propose an end-to-end deep learning framework for semantic segmentation of individual teeth as well as the gingiva from point clouds representing IOS. By introducing a non-uniform resampling technique, our proposed model is trained and deployed on the highest available spatial resolution where it learns the local fine details along with the global coarse structure of IOS. Furthermore, the point-wise cross-entropy loss for semantic segmentation of a point cloud is an ill-posed problem, since the relative geometrical structures between the instances (e.g. the teeth) are not formulated. By training a secondary simple network as a discriminator in an adversarial setting and penalizing unrealistic arrangements of assigned labels to the teeth on the dental arch, we improve the segmentation results considerably. Hence, a heavy post-processing stage for relational and dependency modeling (e.g. iterative energy minimization of a constructed graph) is not required anymore. Our experiments show that the proposed approach improves the performance of our baseline network and outperforms the state-of-the-art networks by achieving 0.94 IOU score.

**Keywords:** Deep learning, 3D point cloud, intra-oral scan, semantic segmentation.

## 1. Introduction

The emergence of digital equipment for extra-oral (e.g. X-ray panoramic, cephalometric and cone beam computed tomography) and intra-oral imaging (e.g. laser or structured light projection scanners) has been a driving force for developing computer-aided design (CAD) systems to analyze the imaging data for highly accurate treatment planning. The purpose of this paper is to explore a segmentation methodology based on deep learning for providing useful clinical information to support better treatment. For supporting an automated clinical workflow in implantology and orthodontic fields, such a CAD system should be able to resolve some fundamental issues of which accurate semantic segmentation of teeth and gingiva (gums) from imaging data is highly desirable. Here, the

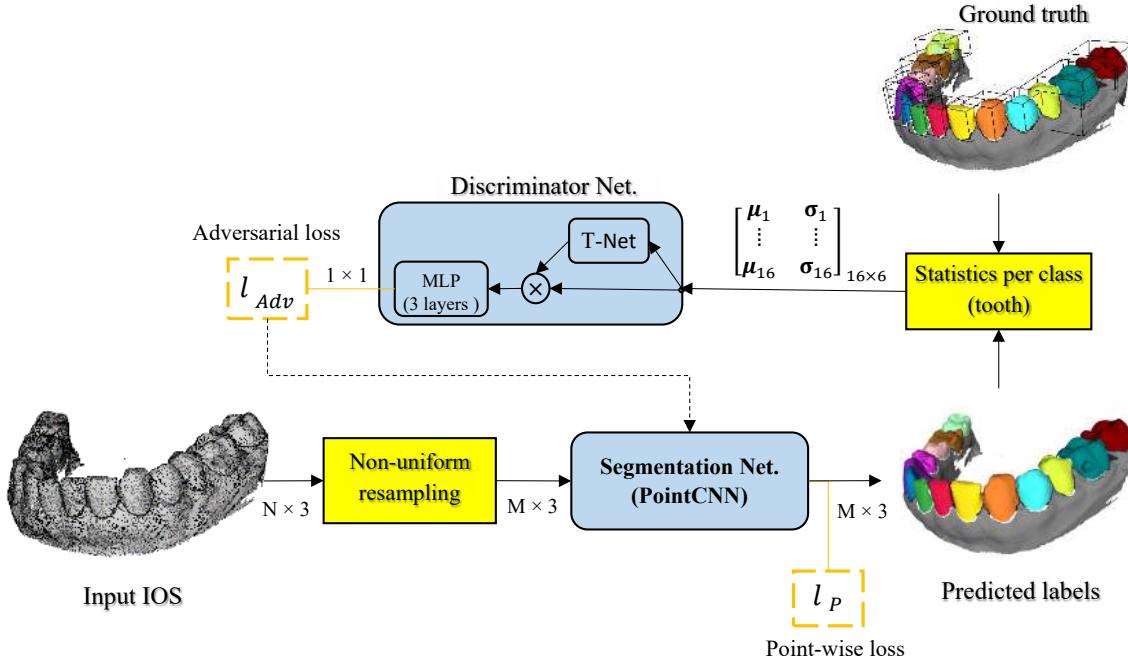


Figure 1: Block diagram of the proposed method in training mode.

semantic segmentation problem for intra-oral scans (IOS) refers to assigning a label, based on the *Fédération Dentaire Internationale* (FDI) standard. In more technical details, this involves labeling all points as belonging to a specific tooth crown or as belonging to gingiva within the recorded IOS point cloud. Each point is represented by a coordinate in the 3D Cartesian coordinate system which is not universal (i.e. the latter can be different between two IOS). The FDI specifies 32 labels for adult dentition, referring to 16 teeth in each upper and lower jaw. In this study, we treat the teeth on the upper and lower jaw in the same way, so that we only employ 16 separate labels to be classified. This changes the problem to finding 16 classes (+1 extra for the gingiva), which facilitates better learning.

To bring artificial intelligence (AI) into modern dentistry, we improve IOS semantic segmentation by means of end-to-end learning of a segmentation model. Building an accurate segmentation model involves two aspects of complexity. Firstly, complexity originates from the dentition (teeth arrangement) and data acquisition. Since the shape of two adjacent tooth crowns (e.g. two molar teeth) may appear to be similar, assigning a correct label demands additional information such as relative position with respect to other teeth on the dental arch. Furthermore, presence of abnormalities in dentition and shape deformation, makes IOS segmentation a challenging task for a segmentation model. An examples of such an abnormality may be lacking teeth (e.g. wisdom teeth). Additional challenges may arise from acquisition issues such as partially missing data (e.g. because of occlusion in scanning), lack of a universal coordinate system, presence of noise, outliers, etc. The interaction of these challenges is important for successfully applying computer vision algorithms.

The second aspect of complexity relates to the 3D geometrical representation of data by a point cloud that is not well suited to the decent deep learning models that are highly performant

on 2D/3D images. Application of such deep learning models (e.g. CNN-based architectures) to point cloud analysis would require three main issues to be addressed. These are: (1) data irregularity,(2) permutation-invariance and (3) resampling-invariance. These issues are discussed briefly below.

**Irregularity** of the point cloud means that the data elements are not organized on a 2D/3D grid, like the data in 2D/3D images. This mainly originates from the pseudorandom nature of recording (sampling) of the external surface of an object, recorded by e.g. a laser scanner. The irregularity results in an ineffective use of convolutional filters for capturing the spatial-local correlation in data (Li et al., 2018), as they work best on organized data.

**Permutation-invariance** refers to the geometrically unordered presentation of a point cloud. If we present a point cloud by a matrix in which each row contains a point, alternating the order of the rows does not change the data semantics, while it does affect the numerical computation in deep learning architectures.

**Resampling-invariance** is a property that means random selection of a sufficiently large subset of the points, preserving the global structure of the object captured by the overall point cloud. The IOS data contains tens of thousands of points. The number of points can vary considerably between two scans, or even between different acquisition runs of the same object. Processing such large-scale and variable-size data is challenging for a deep learning model. Hardware limitations (e.g. memory of the GPU) and working with fixed-rank matrices require a resampling stage. However, a naive resampling approach can invoke the loss of important information and is highly application-dependent.

Since 2016, several studies have investigated point cloud analysis by artificial neural networks (ANNs) for object classification/segmentation tasks. *PointNet* (Qi et al., 2017) and *DeepSets* (Ravanbakhsh et al., 2016) are two pioneering works from recent years, based on the multi-layer perceptron (MLP) network, recently followed by other researchers (Le and Duan, 2018; Li et al., 2018). Available deep learning models include some inventive techniques for the joint handling of the first two mentioned issues (i.e. irregularity and permutation invariance), while still addressing the third issue by applying a uniform resampling for fixing the number of points. Although such an approach is sufficient for many applications like object classification (e.g. classifying the chairs vs. tables), it does not preserve the finer details of data which is important for the segmentation tasks (e.g. classifying a point close to the borderline of a tooth and gingiva). This last issue if not addressed, causes significant performance loss in semantic segmentation tasks.

In this paper, we propose an end-to-end learning framework for IOS segmentation based on recent point cloud deep learning models. Our contribution is threefold.

1. To the best of our knowledge, this is the first end-to-end learning study, proposed for IOS point cloud segmentation.
2. We propose a unique non-uniform resampling mechanism, combined with a compatible loss function, for training and deploying a deep network. The non-uniform resampling facilitates the training and deployment of the network on a fixed-size resampled point cloud which contains different levels of spatial resolution, involving both local, fine details and the global shape structure.
3. In addition to a point-wise classification loss, we employ an adversarial loss for empowering the segmentation network to learn the realistic layout of the labeling space and improving the

classification of points by involving the high-level semantics and preserving the valid arrangement of the teeth labels on the dental arch. In contrast to the existing similar approaches, the discriminator network is applied only to the statistics which are computed from the spatial distributions of labels and the predictions. Consequently, only a shallow network is employed as discriminator that facilitates the training.

## 2. Related work

Related literature has been divided into two parts: conventional IOS segmentation methods and available deep learning solutions for geometric point cloud IOS analysis.

**Conventional IOS segmentation approaches:** The existing literature on IOS segmentation is extensive and based on conventional computer graphic/vision algorithms. Among the proposed methods, one generic approach is first projecting the 3D IOS mesh on one or multiple 2D plane(s) and then applying standard computer vision algorithms. Afterwards, the processed data is projected back into the 3D space. For example, Kondo *et al.* (Kondo et al., 2004) proposes gradient orientation analysis and Wongwaen *et al.* (Wongwaen and Sinthanayothin, 2010) applies a boundary analysis on a 2D projected panoramic depth images for finding teeth boundaries. Most of other studies are based on curvature analysis (Yuan et al., 2010; Kumar et al., 2011; Yaqi and Zhongke, 2010; Yamany and El-Bialy, 1999; Zhao et al., 2006), *fast marching watersheds* (Li et al., 2007), *morphological operations* (Zhao et al., 2006), 2D (Grzegorzek et al., 2010) and 3D (Kronfeld et al., 2010) active contour (snake) analysis and tooth-target harmonic fields (Zou et al., 2015) for segmenting the teeth and gingiva. Some other works follow a semi-automatic approach by manually setting a threshold(Kumar et al., 2011), picking some representative points (Yamany and El-Bialy, 1999), or interactively involve a human operator for the analysis (Yaqi and Zhongke, 2010; Zhao et al., 2006). Such a method is always limited by having to find the best handcrafted features, the manual tuning of several parameters, and also the inherent limitation of handcrafted CAD systems.

**Deep learning approaches:** The available deep learning approaches for structured learning on geometric point clouds can be roughly categorized into four types: *feature-based deep neural networks (DNNs)*, *volumetric*, *2D projection* and *point cloud* methods. *Feature-based DNNs* first extract a set of standard shape features (e.g. based on computer graphic algorithms) and then apply a neural network (e.g. a CNN) for feature classification (Guo et al., 2015; Fang et al., 2015). The performance of this approach is limited to the discriminating properties of the handcrafted features (Qi et al., 2017). The *volumetric* approach, first voxelizes the shape and then applies 3D CNN models on the quantized shape into a 3D grid space (Wu et al., 2015; Qi et al., 2016). As expected, the spatial quantization constrains such a method's performance, especially when fine, high-frequency details need to be preserved in shape curvatures for accurate prediction. The *2D projection* approach first renders the 3D data into one/multiple 2D plane(s) and then applies the 2D convolution operator for the 2D-image pixel classification and then the processed data is projected back into the 3D data (Kalogerakis et al., 2017). *Point cloud* deep learning models work directly with raw point clouds (Qi et al., 2017; Ravanbakhsh et al., 2016; Li et al., 2018; Le and Duan, 2018). Each point has some attributes, mainly their 3D coordinates and sometimes other attributes like the normal of a surface they may represent, color, etc. Currently, point cloud deep learning models are a very active research track. This last approach does not suffer from some shortcomings that occur when using handcrafted features, quantization errors or high processing demands, as is the case with earlier mentioned approaches.

In this paper, we have setup our methodology and experiments for teeth semantic segmentation based on the PointCNN model (Li et al., 2018). The PointCNN model is based on a  $\chi$ -Conv operator, which weighs and permutes the input points and their corresponding features, prior to processing them by a typical convolution operator. The field of view of each  $\chi$ -Conv operator consists of a fixed set of k-nearest neighbour (KNN) points. The outcome of the  $\chi$ -Conv operation is the aggregation and projection of KNN point features into a representative set of points, after which a typical convolution is applied to them. The PointCNN has a lower amount of parameters and has been shown to be effective for learning local correlations from point cloud data (Li et al., 2018). This is beneficial, because it is less prone to severe overfitting on a small dataset.

### 3. Method

The block diagram for our proposed method is shown in Figure 1. In the following section, we will discuss our proposed framework in three parts: preprocessing and data augmentation, non-uniform resampling, and model architecture.

#### 3.1. Data augmentation

The training data are augmented by random 3D rotations, point ordering permutations, adding artificial noise (in the form of jittering) to the positions of each point, and instance dropouts. Here, the dropout of instances means randomly removing all points that belong to a specific tooth from the point cloud in each batch of the training data. This helps the network to learn the labels that may be lacking and do not occur in the training set. The only preprocessing that is applied to the input point cloud is normalization of coordinate information within a scan to have a zero mean and unit variance.

#### 3.2. Non-uniform resampling

Because of the mentioned resampling-invariance property of the point cloud, training a deep learning model on whole set of points of an IOS point would lead to potential issues, such as the variable-rank matrices (the number of points in our IOS datasets may vary in amount between [100k, 310k]) as well as the hardware limitations (such as available memory) for processing of the large-scale point cloud. Applying a patch-classification technique which is common for large-size 2D/3D images, would degrade the quality of results because the extracted patches (i.e. a local subset of points) lack global-structure contents. Furthermore, it would also miss the existing strong dependency between the label of each point and its location in the point cloud. Unfortunately, as we already mentioned, the alternative solution based on uniform resampling does not lead to an accurate analysis of data at its highest available resolution. Recently, various non-uniform resampling methods have been proposed by means of optimization of different metrics that preserve high-frequency contents (Chen et al., 2018; Huang et al., 2013) or local directional density (Skrodzki et al., 2018). However, the effectiveness of using such data abstraction methods on the performance of a deep network cannot be easily established and is in contrast with our interest in designing an end-to-end learning scheme that works directly on the raw data. It is preferable to have such an abstraction of information be performed by the network itself with respect to its objective function. Our proposed non-uniform resampling method is based on the Monte Carlo sampling technique and results in a locally-dense and globally-sparse subset of points for training the deep learning model.

We now state the problem more formally. We assume a matrix representation for the point cloud ( $X = [x_1, x_2, \dots, x_N]$ ) with  $N$  points of which each point has  $D$  attributes. The point  $x_i \in \mathbb{R}^D$  and the point cloud  $X \in \mathbb{R}^{N \times D}$ , where  $D = 3$  for the 3D geometric points. By introducing a radial basis function (RBF), denoted by  $\mathcal{K}$ , which is positioned on a randomly chosen point ( $x_{fovea} \in X$ ), the geometrical similarity (spatial distance) to the point  $x_{fovea}$  can be measured with a weighted distance metric, as specified in Eq.(1). In accordance with the *foveation* as defined in the work of Ciresan *et al.* (Ciresan et al., 2012), we call this point the *fovea*. The RBF kernel is specified by:

$$\mathcal{K}(x_i, x_{fovea}) = \exp\left(-\frac{\|x_i - x_{fovea}\|^2}{2\sigma^2}\right), \quad (1)$$

where  $\sigma$  is a free parameter that controls the bandwidth (*compactness*) of the kernel. By resampling, we aim to choose a subset  $Y$  out of  $X$  with  $M$  points ( $M < N$ ) that has a dense sampling around the fovea and a sparse sampling for farther locations. According to Monte Carlo sampling, by randomly drawing (with replacement) a point  $x_i$  from the set  $X$ , we accept to insert such a point into the subset  $Y$ , only if  $\mathcal{K}(x_i, x_{fovea}) > r_\delta$  is satisfied, otherwise it is rejected. The variable  $r_\delta$  is a random number from a uniform distribution within the unity interval according to the Monte Carlo technique. This process continues until  $M - 1$  unique points are accepted. Algorithm 1 in the Appendix shows these steps in detail. Hence, the resampled subset  $Y$  has  $M$  total points at different levels of granularity (see Figure 2). By random selection of the fovea in every training batch, the model trains on the whole point cloud in its highest available resolution with a fixed number of points. It worths to mention that as the point cloud is normalized to have variance of unity, the uniform-resampling and patch sampling both can be considered as two extreme cases of our proposed algorithm by setting  $\sigma \gg 1$  and  $\sigma \ll 1$ , respectively.

### 3.3. Model architecture

Our proposed model includes two networks: the *segmentation* network ( $\mathcal{S}$ ) and the *discriminator* network ( $\mathcal{D}$ ). The PointCNN (Li et al., 2018) architecture is used for implementing the  $\mathcal{S}$  network. The inputs to the  $S$  network are the resampled points and its output is a 17-element vector for each point, which represents the class probability.

**Weighted point-wise cross entropy loss:** Training the segmentation network by computing an equally weighted loss for each point in the input non-uniform resampled data is not efficient. Since the resampled point set contains various levels of granularity, equally penalizing the output errors for dense and sparse regions prevents the model from optimally adapting its convolutional kernels to capture the fine-detailed content in the data, as the error on sparse points increases relatively equally. Figure 2 shows the uncertainty values for each point, predicted by the network with an equally weighted loss function. As expected, the sparse regions with a lower sampling rate yield a high uncertainty during the learning process because the missing context makes it difficult for the model to perform as accurately as it performs in dense regions. For optimizing the performance of the learning algorithm, we have to trade-off the preservation of the sparse points (which contain the global dental arch structure) and learning of the fine-curvature in point cloud data by parameter tuning. To do so, we apply different weights per point which are computed with the distance metric of the RBF kernel (Eq.(1)). By assuming the posterior probability vector ( $\mathbf{P}_i$ ) for point  $i$ , which is computed at the output softmax layer of the segmentation network with the transfer function of  $\mathcal{S}$

and its parameters  $\theta_{\mathcal{S}}$ , the weighted loss value ( $\mathcal{L}_p$ ) for each point  $i$  is formulated by:

$$\begin{aligned} \mathbf{P}_i = [p_{i1}, \dots, p_{iL}] &= \mathcal{S}(x_i, \theta_{\mathcal{S}}), \quad \text{where} \quad \sum_{j=1}^L p_{ij} = 1 \quad \text{with} \quad 0 \leq p_{ij} \leq 1, \\ \mathcal{L}_p &= - \sum_{i=1}^M w_i \cdot \sum_{j=1}^L y_i \cdot \log(p_{ij}), \quad \text{and} \quad w_i = \mathcal{K}(x_i, x_{fovea}). \end{aligned} \quad (2)$$

Here, the  $y_i$  represents the one-hot encoded target label for the point  $i$  with  $x_i$  in 3D coordinates. In our experiments,  $L = 17$  and  $M = 3 \cdot 10^4$  denote the number of labels and the number of resampled points, respectively.

**Adversarial loss** Training the segmentation network only by applying a standard pixel-wise (voxel or point-wise) cross-entropy loss function has an important shortcoming. The label of each point in the cloud has a high dependency to the label of its adjacent points. For example, if a point belongs to an incisor tooth, its adjacent points can only belong to the same or another incisor, a canine tooth or to the gingiva, but certainly not belong to a molar tooth. Although such a strong structural constraint exists in the data, it is ignored when the optimization problem is only formulated by Eq. (2). As discussed in (Ghafoorian et al., 2018), the semantic segmentation is inherently not a pixel-based (point-wise) classification problem, hence such a formulation is ill-posed. For improving the higher-level semantic consistencies, Luc et al. (Luc et al., 2016) employed an adversarial training in addition to a supervised training of the segmentation network. According to such an approach, a discriminator network provides supervisory signal (feedback) to the segmentation network based on differences between distributions of labels and predictions. Such an effective mechanism was later followed for medical image analysis (Dai et al., 2018; Huo et al., 2018; Kohl et al., 2017; Moeskops et al., 2017; Xue et al., 2018; Yang et al., 2017).

In (Ghafoorian et al., 2018), the authors use a discriminator network to discriminate between the generated labels from a segmentation network and the ground truth labels. Furthermore, they propose using an embedded loss (distance between the features of the hidden layer in the discriminator network) for stability of the training. For point-cloud semantic segmentation, we follow a similar approach, but instead of a heavy training of the discriminator directly on the input space (point cloud and labels) and defining an embedded loss, we first compute two statistical parameters from both the predicted labels and real labels. Afterwards, by training a shallow MLP network as a discriminator, we facilitate the segmentation network's ability to produce a more realistic prediction. The statistics that we used simply consist of the mean and variance of the coordinates of all points with the same label, as given by the segmentation network, which leads to:

$$\hat{\mu}_j = \sum_{i=1}^M p_{ij} \cdot x_i \quad \text{and} \quad \hat{\sigma}_j^2 = \sum_{i=1}^M p_{ij} \cdot (x_i - \hat{\mu}_j)^2, \quad j = 1, 2, \dots, L-1 \quad (3)$$

$$\hat{\mathbf{u}}_{(L-1) \times 6} = [\hat{\mu}_1, \hat{\sigma}_1^2 \parallel \hat{\mu}_2, \hat{\sigma}_2^2 \parallel \dots \parallel \hat{\mu}_{L-1}, \hat{\sigma}_{L-1}^2]. \quad (4)$$

Here, the  $\parallel$  denotes a vertical vector concatenation (stacking). As mentioned earlier,  $L$  denotes the number of labels in the data. The stacked feature set ( $\hat{\mathbf{u}}$ ) represents a *soft* computation of the central positions of teeth and their variance (i.e. their soft bounding boxes) in the 3D space, according to the predicted labels ( $p_{ij}$ ). The statistical mean and variance are computed only for  $L-1$  classes of

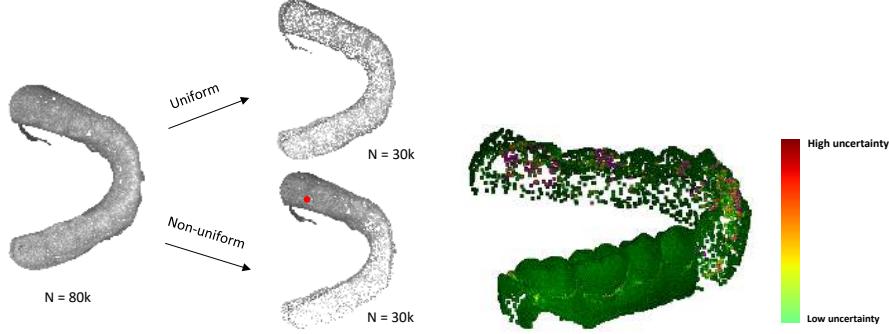


Figure 2: Example of uniform vs. non-uniform resampling (left). The fovea is shown by a red dot. The uncertainty of the prediction for dense and sparse regions (right).

teeth. In computing the high-level semantic features (statistics), we ignore the gingiva class, since its point cloud is almost spread across the whole input space and the applied non-uniform resampling stage alters its resulting statistics severely across training batches. By replacing the  $p_{ij}$  values in Eq. 3 with the one-hot encoded values of the ground truth labels ( $y_i$ ), the counterpart feature-set of  $\hat{u}$ , denoted by  $u$ , is obtained. For any absent label in the point cloud, we simply insert a vector consisting of zeros instead. The feature set  $u$  represents a *realistic* statistical measurement of the labeled data.

The discriminator network ( $\mathcal{D}$ ) aims to discriminate between feature set  $u$  and  $\hat{u}$ . The network consists of two cascaded parts. The first part estimates an affine transformation and is applied to a input 96-element input vector. The second part consists of 3-layer MLP network which maps the transformed input vector into a scalar value by a sigmoidal activation function at its output node. In effect, the network tries to produce the scalar 1 at its output if the network is applied on  $u$ , while the scalar 0 should be produced if the network is applied on  $\hat{u}$ . The architecture of the first part of the network is identical to what is proposed in the PointNet model (Qi et al., 2017), called a *T-Net*. More details about the T-Net can be found in (Qi et al., 2017). In an adversarial setting for training the network  $D$  and network  $S$ , the discriminator loss ( $\mathcal{L}_{\mathcal{D}}$ ) for the network  $\mathcal{D}$  with parameters  $\theta_{\mathcal{D}}$  and an adversarial loss for the network  $\mathcal{S}$ , can be written as:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(u, \hat{u}; \theta_{\mathcal{D}}, \theta_{\mathcal{S}}) &= \mathbb{E}_u [\log \mathcal{D}(u)] + \mathbb{E}_{\hat{u}} [\log(1 - \mathcal{D}(\hat{u}))], \\ \mathcal{L}_{Adv}(\hat{u}; \theta_{\mathcal{D}}, \theta_{\mathcal{S}}) &= \mathbb{E}_{\hat{u}} [\log \mathcal{D}(\hat{u})].\end{aligned}\quad (5)$$

Hence, the total loss for the segmentation network is a contribution of the losses  $\mathcal{L}_p$  in Eq.(2) and  $\mathcal{L}_{Adv}$  in Eq. (5). To avoid the need for manual hyper-parameter tuning for the contribution weights ( $\lambda$ ) between two loss terms, we follow the work by Kendall *et al.* (Kendall et al., 2017) and involve *adaptive loss weighting*. After initializing  $\lambda = [\lambda_1, \lambda_2]$  with a vector of ones, we add the regularization term  $\mathcal{R}(\lambda)$  to the total loss function for the segmentation network ( $\mathcal{S}$ ), giving:

$$\mathcal{L}_{\mathcal{S}} = \frac{1}{\lambda_1^2} \cdot \mathcal{L}_p + \frac{1}{\lambda_2^2} \cdot \mathcal{L}_{Adv} + \mathcal{R}(\lambda), \quad \text{where } \mathcal{R}(\lambda) = \lambda_1^2 \cdot \lambda_2^2. \quad (6)$$

**Inference on the whole point cloud:** Since the segmentation network is trained on non-uniformly resampled data, for prediction on the whole point cloud we need to extract several subsets of points

according to the non-uniform resampling algorithm. Afterwards, the prediction of all points in the original point cloud is obtained by aggregating all the estimated labels for the extracted subsets. The pseudocode of Algorithm 2 in the Appendix describes this procedure in detail.

## 4. Experiment and Results

### 4.1. Data

Our dataset consists of 120 optical scans of dentitions from 60 adults subjects, each containing one upper and one lower jaw scan. The dataset includes scans from healthy dentitions and a variety of abnormalities among subjects. The optical scan data was recorded by a 3Shape d500 optical scanner (3Shape AS, Copenhagen, Denmark). On average, an IOS contains 180k points (varying in range between [100k, 310k]). All optical scans were manually segmented and their respective points were categorized into one of the 32+1 classes by a dental professional and reviewed and adjusted by one dental expert (DAM) with Meshmixer 3.4 (Autodesk Inc, San Rafael CA, USA). Labeling of the tooth categories was performed according to the international tooth numbering standard (FDI). Segmentation of each optical scan took 45 minutes on average, which shows its intensive laborious task for a human.

### 4.2. Experimental setup

The performance of the model is evaluated by fivefold cross-validation and the results are compared making use of the average Jaccard Index, also known as intersection over union (IoU). On top of the IoU, we report the precision and recall for our multi-class segmentation problem. For computing the precision and recall, each class is treated individually (one-versus-all), as a binary problem and finally the the average scores are reported. Our experiments are partitioned into three parts: (1) benchmarking the performance of the PointCNN in comparison with two other state-of-the-art deep learning models capable of IOS segmentation. These models include PointNet (Qi et al., 2017) and PointGrid (Le and Duan, 2018); (2) Evaluating the impact of applying the non-uniform resampling versus using naive uniform resampling. For the purpose of fair comparison, the number of resampled points are kept identical ( $M = 30k$ ); (3) Evaluating the effectiveness of involving the adversarial loss.

All models are trained utilizing stochastic gradient descent and the Adam learning adaptation technique for 1000 epochs with batch size of one. The initial learning rate is equal to  $5e-3$ , which decreases each  $20K$  iterations by a factor of 0.9. We empirically adjust the free parameter of the resampling kernel (see Eq.(1)) to 0.4 (i.e.  $\sigma = 0.4$ ). Since the point cloud is normalized to have a unit variance, we have found that the resampled point cloud by such a chosen setting of  $\sigma$  would encompass at least two teeth in its dense region.

### 4.3. Results

Table 1 depicts the obtained results from our different experimental setups. Figure 3 in the Appendix shows visualizations of a number of exemplary results from our proposed model. As we can observe from Table 1, the PointCNN performs better than two other state-of-the-art models when a naive uniform resampling is applied. This is mostly because of the inclusion of the spatial-correlation information by the  $\chi - Conv$  operator in the PointCNN and its lower amount of parameters, which is less prone to overfitting. The PointGrid which samples points inside a predefined grid utilizes

Table 1: Performance of the proposed model within different experimental setups in comparison with state-of-the-art models.

Method			Metric			Exec.time (sec.)
Network Arch.	Non-uniform	Adv. setting	IoU	Precision	Recall	
<b>PointNet (Qi et al., 2017)</b>	-	-	.76	.73	.65	<b>0.19</b>
<b>PointGrid (Le and Duan, 2018)</b>	-	-	.80	.75	.70	0.88
<b>PointCNN (Li et al., 2018)</b>	-	-	.88	.87	.83	0.66
<b>Proposed (I)</b>	✓	-	.91	.90	.87	6.86
<b>Proposed (II)</b>	-	✓	.91	.91	.89	0.66
<b>Proposed (III)</b>	✓	✓	<b>.94</b>	<b>.93</b>	<b>.90</b>	6.86

convolutional operators, but its performance is still limited to the spatial resolution of the spatial quantization grid. The PointNet performance is also constrained, as it omits processing of spatial correlations in the point cloud. With the choice of basing of our method on PointCNN, we show the effectiveness of applying non-uniform resampling and the adversarial loss. The last two techniques improve the results. Finally, incorporating both techniques simultaneously, the highest performance is achieved.

## 5. Discussion and conclusion

In this paper, we propose an end-to-end learning approach for semantic segmentation of teeth and gingiva from point clouds derived from IOS data. Our segmentation network is based on PointCNN, which has been proposed for point cloud classification/segmentation tasks. For analysis of point clouds in their original spatial resolution (resulting in predictions for all points), we propose a non-uniform resampling mechanism and a compatible loss weighting, based on foveation and Monte Carlo sampling. This resampling approach includes both local, fine-detail information and the sparse global structure of data, which is essential for an accurate prediction of each individual point in absence of a universal coordinate system. Furthermore, by involving the high-level data semantics, through training a discriminator network for learning the realistic layout of labels in data, the results are improved. As a consequence, a heavy post-processing stage (e.g. applying conditional random fields (CRF) on a constructed graph) is not required for incorporating dependencies and locality constraints into the model. By computing the statistics (mean and variance) from spatial distributions of labels and their predictions and feeding them into the discriminator, the adversarial training of the segmentation network is facilitated since for processing such an abstract data only a shallow network can be employed as discriminator. Here, computing the mean and variance of labels and the predictions can be considered generic enough that does not violate the end-to-end learning scheme of the method as using such statistics (operations) is common even within a CNN (e.g. batch normalization operation).

## References

- Siheng Chen, Dong Tian, Chen Feng, Anthony Vetro, and Jelena Kovačević. Fast resampling of three-dimensional point clouds via graphs. *IEEE Transactions on Signal Processing*, 66(3):666–681, 2018.

- Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.
- Wei Dai, Nanqing Dong, Zeya Wang, Xiaodan Liang, Hao Zhang, and Eric P Xing. Scan: Structure correcting adversarial network for organ segmentation in chest x-rays. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 263–273. Springer, 2018.
- Yi Fang, Jin Xie, Guoxian Dai, Meng Wang, Fan Zhu, Tiantian Xu, and Edward Wong. 3d deep shape descriptor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2319–2328, 2015.
- Mohsen Ghafoorian, Cedric Nugteren, Nóra Baka, Olaf Booij, and Michael Hofmann. El-gan: Embedding loss driven generative adversarial networks for lane detection. *arXiv preprint arXiv:1806.05525*, 2018.
- Marcin Grzegorzek, Marina Trierscheid, Dimitri Papoutsis, and Dietrich Paulus. A multi-stage approach for 3d teeth segmentation from dentition surfaces. In *International Conference on Image and Signal Processing*, pages 521–530. Springer, 2010.
- Kan Guo, Dongqing Zou, and Xiaowu Chen. 3d mesh labeling via deep convolutional neural networks. *ACM Transactions on Graphics (TOG)*, 35(1):3, 2015.
- Hui Huang, Shihao Wu, Minglun Gong, Daniel Cohen-Or, Uri Ascher, and Hao Richard Zhang. Edge-aware point set resampling. *ACM Transactions on Graphics (TOG)*, 32(1):9, 2013.
- Yuankai Huo, Zhoubing Xu, Shunxing Bao, Camilo Bermudez, Andrew J Plassard, Jiaqi Liu, Yuang Yao, Albert Assad, Richard G Abramson, and Bennett A Landman. Splenomegaly segmentation using global convolutional kernels and conditional generative adversarial networks. In *Medical Imaging 2018: Image Processing*, volume 10574, page 1057409. International Society for Optics and Photonics, 2018.
- Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, and Siddhartha Chaudhuri. 3d shape segmentation with projective convolutional networks. In *Proc. CVPR*, page 8, 2017.
- Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv preprint arXiv:1705.07115*, 3, 2017.
- Simon Kohl, David Bonekamp, Heinz-Peter Schlemmer, Kaneschka Yaqubi, Markus Hohenfellner, Boris Hadischik, Jan-Philipp Radtke, and Klaus Maier-Hein. Adversarial networks for the detection of aggressive prostate cancer. *arXiv preprint arXiv:1702.08014*, 2017.
- Toshiaki Kondo, SH Ong, and Kelvin WC Foong. Tooth segmentation of dental study models using range images. *IEEE Transactions on medical imaging*, 23(3):350–362, 2004.
- Thomas Kronfeld, David Brunner, and Guido Brunnett. Snake-based segmentation of teeth from virtual dental casts. *Computer-Aided Design and Applications*, 7(2):221–233, 2010.

- Yokesh Kumar, Ravi Janardan, Brent Larson, and Joe Moon. Improved segmentation of teeth in dental models. *Computer-Aided Design and Applications*, 8(2):211–224, 2011.
- Truc Le and Ye Duan. Pointgrid: A deep network for 3d shape understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9204–9214, 2018.
- Yangyan Li, Rui Bu, Mingchao Sun, and Baoquan Chen. Pointcnn. *arXiv preprint arXiv:1801.07791*, 2018.
- Zhanli Li, Xiaojuan Ning, and Zengbo Wang. A fast segmentation method for stl teeth model. In *Complex Medical Engineering, 2007. CME 2007. IEEE/ICME International Conference on*, pages 163–166. IEEE, 2007.
- Pauline Luc, Camille Couprise, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.
- Pim Moeskops, Mitko Veta, Maxime W Lafarge, Koen AJ Eppenhof, and Josien PW Pluim. Adversarial training and dilated convolutions for brain mri segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 56–64. Springer, 2017.
- Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016.
- Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 1(2):4, 2017.
- Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Deep learning with sets and point clouds. *arXiv preprint arXiv:1611.04500*, 2016.
- Martin Skrodzki, Johanna Jansen, and Konrad Polthier. Directional density measure to intrinsically estimate and counteract non-uniformity in point clouds. *Computer Aided Geometric Design*, 2018.
- Nonlapas Wongwaen and Chanjira Sinthanayothin. Computerized algorithm for 3d teeth segmentation. In *Electronics and Information Engineering (ICEIE), 2010 International Conference On*, volume 1, pages V1–277. IEEE, 2010.
- Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.
- Yuan Xue, Tao Xu, Han Zhang, L Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scale l1 loss for medical image segmentation. *Neuroinformatics*, 16(3-4):383–392, 2018.
- Sameh M Yamany and Ahmed M El-Bialy. Efficient free-form surface representation with application in orthodontics. In *Three-Dimensional Image Capture and Applications II*, volume 3640, pages 115–125. International Society for Optics and Photonics, 1999.

Dong Yang, Daguang Xu, S Kevin Zhou, Bogdan Georgescu, Mingqing Chen, Sasa Grbic, Dimitris Metaxas, and Dorin Comaniciu. Automatic liver segmentation using an adversarial image-to-image network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 507–515. Springer, 2017.

Ma Yaqi and Li Zhongke. Computer aided orthodontics treatment by virtual segmentation and adjustment. In *Image Analysis and Signal Processing (IASP), 2010 International Conference on*, pages 336–339. IEEE, 2010.

Tianran Yuan, Wenhe Liao, Ning Dai, Xiaosheng Cheng, and Qing Yu. Single-tooth modeling for 3d dental model. *Journal of Biomedical Imaging*, 2010:9, 2010.

Mingxi Zhao, Lizhuang Ma, Wuzheng Tan, and Dongdong Nie. Interactive tooth segmentation of dental models. In *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pages 654–657. IEEE, 2006.

Bei-ji Zou, Shi-jian Liu, Sheng-hui Liao, Xi Ding, and Ye Liang. Interactive tooth partition of dental mesh base on tooth-target harmonic field. *Computers in biology and medicine*, 56:132–144, 2015.

## Appendix

---

**Algorithm 1:** Non-uniform resampling

---

**input :** point cloud  
**output:** non-uniform resampled point cloud

```

 $X \leftarrow \{x_1, x_2, \dots, x_N\};$  // whole point cloud
 $Y \leftarrow \emptyset;$  // initialized empty set
 $x_f \leftarrow x \sim X;$  // randomly draw one sample as fovea
Function  $\mathcal{R}(x_f, X):$ 
  while  $|Y| < M;$  // check the size of Y
  do
     $x_i \leftarrow x \sim X;$  // randomly draw sample
     $r_\delta \sim \text{uniform}(0, 1);$  // draw a random value
    if  $\mathcal{K}(x_i, x_f) \geq r_\delta;$  // The RBF kernel Eq. 1
      then
        if  $x_i \notin Y$  then
           $Y \leftarrow x_i \cup Y;$  // insert to the subset
        end
      end
    end
  end
return  $Y$ 

```

---



---

**Algorithm 2:** Inference on the whole point cloud

---

**input :** point cloud  
**output:** predicted label per point

```

 $X \leftarrow \{x_1, x_2, \dots, x_N\};$  // whole point cloud
Function  $Inference(X):$ 
   $U \leftarrow \emptyset;$  // initialized an empty set
   $P_X \leftarrow 0_{N \times 17};$  // initialized probability vectors
  while  $|U| < |X|$  do
     $x_f \sim \{x \in X \mid x \notin U\};$  // select fovea out of the unprocessed points
     $Y \leftarrow \mathcal{R}(x_f, X);$  // non-uniform resampling (Algorithm 1)
     $P_Y = \mathcal{S}(Y, \theta_S);$  // prediction of  $\mathcal{S}$  Net.
     $\{x_i\} \leftarrow \{x \in Y \mid \mathcal{K}(x_f, x) < \sigma\};$  // only dense region is valid
     $P_X(x_i) = P_X(x_i) + P_Y(x_i);$  // Aggregate the probabilities
     $U \leftarrow \{x_i\} \cup \{U\};$  // Mark as processed
  end
return  $\text{argmax}(P_X);$  // labels on whole point cloud

```

---

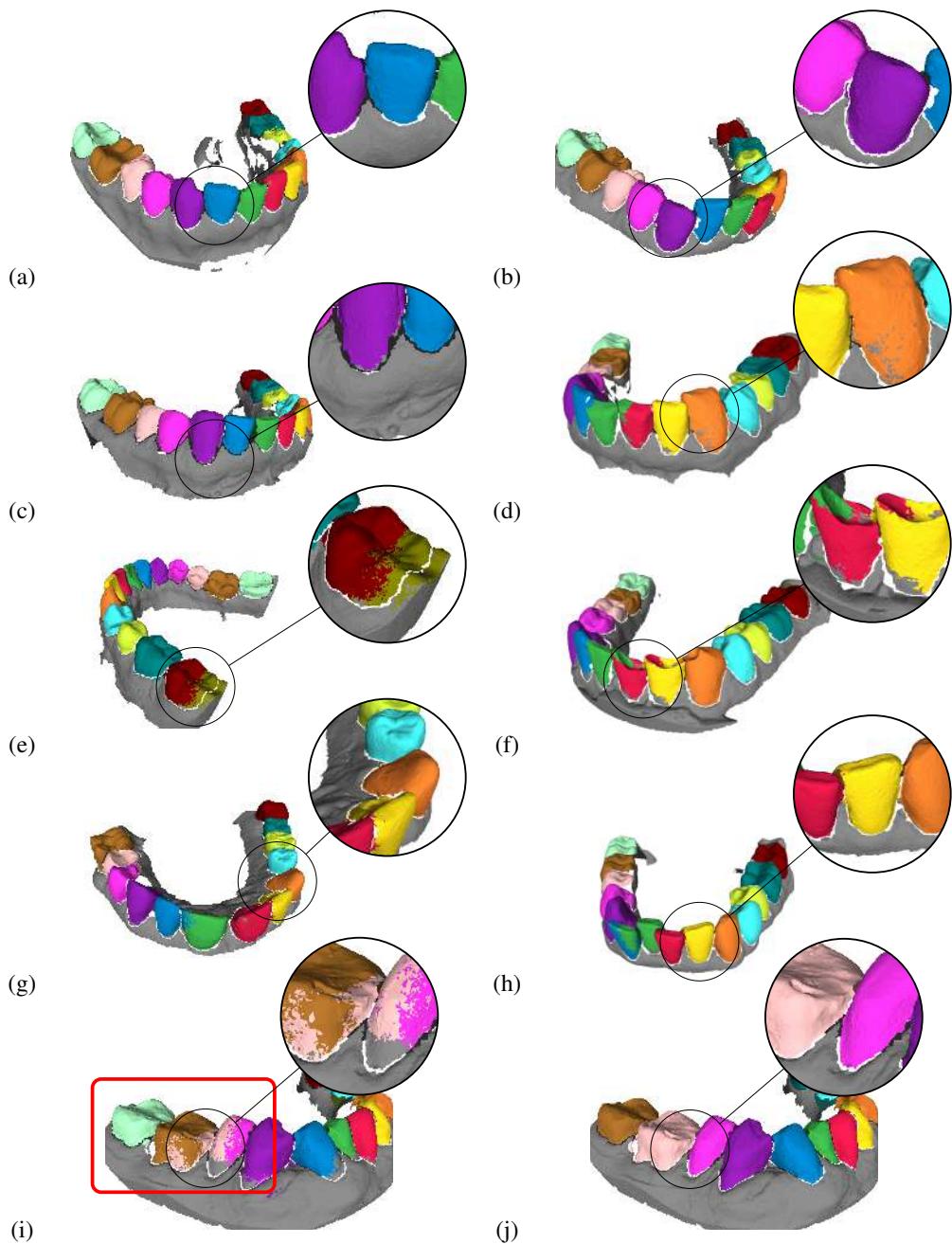


Figure 3: (a-h) Examples of segmentation by our proposed method. (i) Example of a failure case when the adversarial loss is not involved in the training of the segmentation network. The assigned label inside the circle is unrealistic (i.e. invalid). Consequently, the model assigned a set of invalid labels to other neighbouring teeth (inside the red rectangle) by their maximum likelihood. (j) Ground truth.

# SPDA: Superpixel-based Data Augmentation for Biomedical Image Segmentation

**Yizhe Zhang<sup>1</sup>**

YZHANG29@ND.EDU

**Lin Yang<sup>1</sup>**

LYANG5@ND.EDU

**Hao Zheng<sup>1</sup>**

HZHENG3@ND.EDU

**Peixian Liang<sup>1</sup>**

PLIANG@ND.EDU

**Colleen Mangold<sup>2</sup>**

CAV154@PSU.EDU

**Raquel G. Loreto<sup>2</sup>**

RAQUELGLORETO@GMAIL.COM

**David P. Hughes<sup>2</sup>**

DHUGHES@PSU.EDU

**Danny Z. Chen<sup>1</sup>**

DCHEN@ND.EDU

<sup>1</sup>*Department of Computer Science and Engineering, University of Notre Dame, USA*

<sup>2</sup>*Department of Entomology and Department of Biology, Center for Infectious Disease Dynamics, Pennsylvania State University, USA*

## Abstract

Supervised training a deep neural network aims to “teach” the network to mimic human visual perception that is represented by image-and-label pairs in the training data. Superpixelized (SP) images are visually perceptible to humans, but a conventionally trained deep learning model often performs poorly when working on SP images. To better mimic human visual perception, we think it is desirable for the deep learning model to be able to perceive not only raw images but also SP images. In this paper, we propose a new superpixel-based data augmentation (SPDA) method for training deep learning models for biomedical image segmentation. Our method applies a superpixel generation scheme to all the original training images to generate superpixelized images. The SP images thus obtained are then jointly used with the original training images to train a deep learning model. Our experiments of SPDA on four biomedical image datasets show that SPDA is effective and can consistently improve the performance of state-of-the-art fully convolutional networks for biomedical image segmentation in 2D and 3D images. Additional studies also demonstrate that SPDA can practically reduce the generalization gap.

## 1. Introduction

Traditional data augmentation methods use a combination of geometric transformations to artificially inflate training data (Perez and Wang, 2017). For each raw training image and its corresponding annotated image, it generates “duplicate” images that are shifted, zoomed in/out, rotated, flipped, and/or distorted. These basic/traditional data augmentation methods are generally applicable to classification problems where the output is a vector and segmentation problems where the output is a segmentation map.

Recently, generative adversarial networks (GANs) have been used for data augmentation (e.g., (Antoniou et al., 2017)). Encouraging the generator to produce realistic looking images (comparing to the original images) is a main consideration when training the generator. A key issue to this consideration is that it does not define/imply what kind of generated images would be use-

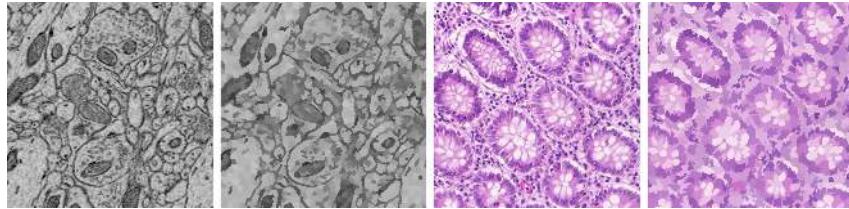


Figure 1: From left to right: An electron micrograph of neuronal structure, its superpixelized image, an H&E stained pathological image of glands, and its superpixelized image. The superpixels preserve the essential objects and their boundaries.

ful/meaningful for data augmentation purpose, and the generator does not necessarily converge to a model version that generates useful new data for training a better segmentation or classification model. (Wang et al., 2018) was proposed to deal with this issue using a task-related classifier for training an image generator. However, the method in (Wang et al., 2018) was designed for classification problems; in segmentation, the distributions of labels are usually much more complicated and it is quite non-trivial to extend the method (Wang et al., 2018) to segmentation tasks.

As an algorithm based (non-learning based) data augmentation technique, mixup (Zhang et al., 2017) was proposed to generate new image samples “between” pairs of training samples for image classification problems. It was motivated based on the principles of Vicinal Risk Minimization (Chapelle et al., 2001) and its experimental results showed promising classification accuracy improvement. In (Eaton-Rosen et al., 2018), it extended the mixup method to medical image segmentation, showing that mixup is also applicable to data augmentation for segmentation problems.

In this paper, we propose a new algorithm-based data augmentation technique that uses superpixels for better training a deep learning model for biomedical image segmentation. Our method is based on a common experience that superpixelized (SP) images are visually perceivable to humans (see Fig. 1), but a conventionally trained deep learning model (trained using only raw images) often performs poorly when working on SP images. This phenomenon implies that a conventionally trained deep learning model may not mimic human visual behaviors well enough. Thus, we think encouraging a deep learning network to be able to perceive not only raw images but also SP images can make it more closely mimic human visual perception. Our method is built on this idea, by adding SP images to the training data for training a deep learning model. Our new superpixel-based data augmentation (SPDA) method can work together with traditional data augmentation methods and be generally applicable to many deep learning based image segmentation models.

A short summary of our SPDA method is as follows. For each raw image, we apply a superpixel generation method (e.g., SLIC (Achanta et al., 2012)) to obtain superpixel cells. Superpixel cells are groups of pixels that are visually similar and spatially connected. For every superpixel cell  $C$ , we compute the average pixel value(s) for all the pixels in  $C$  and assign the computed average value(s) to all the pixels in  $C$ . In this way, we effectively remove very local image details and emphasize more on the overall colors, shapes, and spatial relations of objects in the image (see Fig. 1). After “superpixelizing” all the raw images in the original training set, we put all the superpixelized (SP) images into the training set together with the original training images for training a deep learning model. Our experiments of SPDA on four biomedical image datasets show that SPDA is effec-

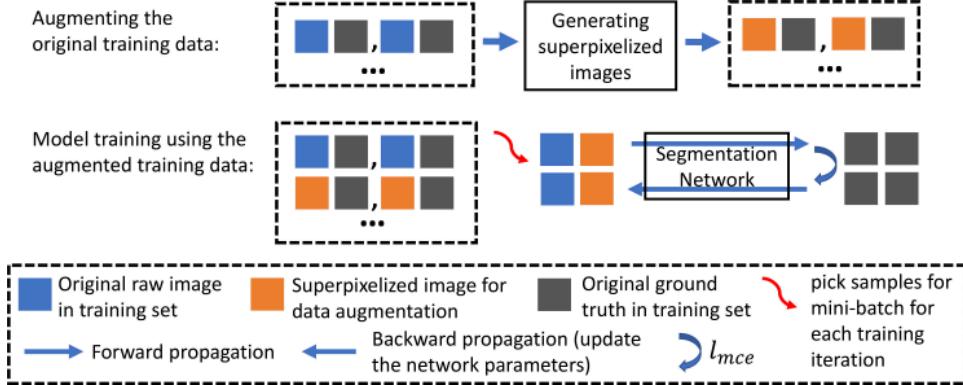


Figure 2: An overview of our SPDA method. During training, we first generate superpixelized images for all the raw images, and then add the SPDA-generated data to the training data for training a segmentation model. Note that the trained network is applied to only raw images during model testing.

tive and can consistently improve the performance of state-of-the-art fully convolutional networks (FCNs) for biomedical image segmentation in 2D and 3D images.

In Section 2, we discuss several technical considerations on generating superpixelized images for data augmentation, and present our exact procedure for generating and using superpixelized images to train deep learning models. In Section 3, we evaluate SPDA using multiple widely used FCNs on four biomedical image segmentation datasets, and show that SPDA consistently yields segmentation performance improvement on these datasets.

## 2. Superpixels for Data Augmentation

First, we give some notation and background of data augmentation. Then, we discuss several technical considerations on using superpixels for data augmentation. Finally, we present the key technical components: (i) What superpixel generation method we choose to use and the logic behind it; (ii) the exact procedure for generating superpixelized images; (iii) the training objective function and algorithm for using SPDA-generated images in deep learning model training. Fig. 2 gives an overview of our SPDA method for model training.

### 2.1. Notation and preliminaries

Given a set of image samples  $X = \{x_1, \dots, x_n\}$  and their corresponding ground truth  $Y = \{y_1, \dots, y_n\}$ , for training a segmentation model (e.g., an FCN)  $f \in F$  that describes the relationship between  $x_i$  and  $y_i$ , the empirical risk is:

$$\frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(x_i), y_i) \quad (1)$$

where  $\mathcal{L}$  is a loss function (e.g., the cross-entropy). Learning the function  $f$  is by minimizing Eq. (1), which is also known as Empirical Risk Minimization.

One could use some proper functions to generate more data based on the original training data pair  $(x_i, y_i)$ . In general, we denote the generated data by  $(x_i^{aug}, y_i^{aug})$ .

When there are multiple ( $k$ ) versions of augmented data for one pair  $(x_i, y_i)$ , the loss with augmented data can be written as:

$$\frac{1}{n} \sum_{i=1}^n (\mathcal{L}(f(x_i), y_i) + \lambda \sum_{j=1}^k \mathcal{L}(f(x_i^{aug_j}), y_i^{aug_j})) \quad (2)$$

where  $\lambda$  is a hyper-parameter that controls the importance of the data augmentation term. Different ways of data augmentation produce different new data, and thus directly affect the learning procedure of  $f$ . As a common practice, flipping, rotation, cropping, etc. are widely used for data augmentation. This type of data augmentation applies geometric transformations ( $g_k$ , for  $k$  different geometric transformations) to both  $x_i$  and  $y_i$ , to generate new pairs of training data. For this type of data augmentation, Eq. (2) can be rewritten as:

$$\frac{1}{n} \sum_{i=1}^n (\mathcal{L}(f(x_i), y_i) + \lambda \sum_{j=1}^k \mathcal{L}(f(g_j(x_i)), g_j(y_i))) \quad (3)$$

Another type of data augmentation makes no change on  $y_i$ , and the only modification/augmentation is on  $x_i$  (e.g., color jittering (Krizhevsky et al., 2012)). For this type of augmentation, Eq. (2) can simply be:

$$\frac{1}{n} \sum_{i=1}^n (\mathcal{L}(f(x_i), y_i) + \lambda \sum_{j=1}^k \mathcal{L}(f(G(x_i)), y_i)) \quad (4)$$

where  $G(\cdot)$  is a label-preserving transformation. Our new SPDA method belongs to this category. We propose to generate superpixlized images (denoted by  $SP(\cdot)$ ) as a type of label-preserving (perception-preserving) transformation for data augmentation.

Below we discuss several technical considerations on using superpixels for data augmentation, the technical details of  $SP(\cdot)$ , and how to use SPDA-generated data for model training.

## 2.2. Technical considerations

In this subsection, we discuss three technical considerations on generating superpixelized images for data augmentation.

(1) Superpixelizing an image  $x$  removes or reduces local image details in  $x$  that might be less relevant to modeling  $P(y|x, \theta_f)$  ( $\theta_f$  denotes the parameters of the segmentation model  $f$ ). A superpixelized image  $SP(x)$  is a simplified version of the original image  $x$ . Letting a deep learning model learn from  $SP(x)$  to predict  $y$  means asking the model to use little or no local (insignificant) pixel value changes and focus more on higher-level semantic information. Since model parameters are shared between predicting  $y$  when given  $x$  and predicting  $y$  when given  $SP(x)$ , modeling  $P(y|SP(x), \theta_f)$  will influence modeling  $P(y|x, \theta_f)$ . As a result, because of the joint modeling of  $P(y|SP(x), \theta_f)$ , the learned function for predicting  $y$  given  $x$  would become more invariant/insensitive to local image noise and small details, and would learn and utilize more higher level image information and representations. Note that all the original training images with all their local image details are still fully kept in the training dataset. Hence, whenever needed, the learning procedure is still able to use any local image details for modeling  $P(y|x, \theta_f)$ .

(2) SPDA provides new image samples that are “close” to the original training samples. Under the principle of Vicinal Risk Minimization or VRM (Chapelle et al., 2001), a vicinity or neighborhood around every training sample is defined or suggested based on human knowledge. Additional samples then can be drawn from this vicinity distribution of the training samples to increase or enlarge the support of the training sample distribution (Zhang et al., 2017). Superpixelized images most of the time are conceptually meaningful to human eyes. It is likely that superpixelized images are also close<sup>1</sup> to their corresponding original image samples in the data space. If this “close neighborhood” property is true, then adding SPDA-generated data to the training data should be helpful to improve the generalization capability of the model, according to VRM (Chapelle et al., 2001). In Appendix A.1, we show that after using a generic dimensionality reduction method (e.g., PCA, t-SNE (Maaten and Hinton, 2008)), one can observe that each superpixelized image is in a close neighborhood of its corresponding original image.

(3) Adding superpixelized images to the training set makes the data distribution of the training set thus resulted closer to the test data distribution or the true data distribution. Superpixelized images form a more general and broader base for the visual conception related to the learning task. Adding superpixelized images to the original training data makes the training data distribution have a more generic base that can potentially better support unseen test images. In Appendix A.2, using variational auto-encoders (VAEs) (Kingma and Welling, 2013) and the Kullback-Leibler divergence (Kullback and Leibler, 1951), we show that the training set with SPDA-generated data is closer to the test set in terms of the overall data distribution.

### 2.3. Choosing a superpixel generation method

Boundary recall and compactness are two key criteria for generation of superpixels. Boundary recall evaluates how well the generated superpixels represent or cover the important object boundaries/contours in an image. Compactness describes how regular and well-organized the superpixels are. Compactness of superpixels tends to constrain superpixels to fit some irregular and subtle object boundaries. In general, one aims to generate superpixels with high boundary recall and high compactness.

For deep learning model training, we aim to generate superpixels with the following properties: (i) good boundary recall, (ii) being compact and pixel-like, and (iii) only pixel values and local image features are used to generate superpixels. Note that many fully convolutional networks work in a bottom-up fashion; superpixels that are generated using global-level information may confuse the training of an FCN model. Hence, we prefer to use superpixel generation method that only utilizes local image information for the pixel grouping process.

SLIC (Achanta et al., 2012) is one of the most widely used methods for generating superpixels. SLIC is fast to compute and can produce good quality superpixels with an option to let the user control the compactness of the generated superpixels. Also, SLIC utilizes only local image information for grouping pixels into superpixels, which is a desired feature by SPDA for training deep learning models. Thus, in our experiments, we use SLIC (Achanta et al., 2012) to generate superpixels for our superpixel-based data augmentation method. The added computational cost for applying SLIC to every training sample is very small comparing to the model training time cost.

---

1. Being close means the distance (e.g., Euclidean distance) between an SPDA-generated image and its corresponding raw image is smaller than the distance between this raw image and any other raw image.

## 2.4. Generating superpixelized images

Suppose the given training set contains  $n$  training samples  $(x_i, y_i), i = 1, 2, \dots, n$ , where  $x_i$  is a raw image and  $y_i$  is its corresponding annotation map. We apply a superpixel generation method (e.g., SLIC (Achanta et al., 2012))  $F(x_i, s)$  to each image  $x_i$  to obtain superpixel cells  $c_j^i, j = 1, 2, \dots, s$ . Each superpixel cell contains a connected set of pixels. Here,  $s$  is part of the input to  $F$  that specifies the desired number of superpixels that  $F$  should produce. We will discuss how to choose the values of  $s$  below. Any two different superpixel cells have zero common elements (pixels). The union of the pixels of all the superpixel cells for  $x_i$  is all the pixels in the image  $x_i$ .

To generate a superpixelized image for  $x_i$ , for each superpixel cell  $c_j^i$ , we compute the mean values of all the pixels in  $c_j^i$  and update the values of all the pixels in  $c_j^i$  using such computed mean values. This step aims to erase low-level pixel variance so that the mid-level and high-level information can be better emphasized by the superpixelized images. We repeat this process for all the superpixel cells of  $x_i$ , and then form a superpixelized image  $SP(x_i, s)$ , where  $s$  indicates that this superpixelized image is generated using  $s$  superpixels. To avoid artificially changing the distribution of annotation (label) maps, the annotation map for  $SP(x_i, s)$  is kept as the original  $y_i$ . Thus, we put  $(SP(x_i, s), y_i)$  into our new training data set generated using  $(x_i, y_i)$ .

The value  $s$  specifies the desired number of superpixels to generate. A small number of superpixels would make a superpixelized image too coarse to represent the essential object structures in the original image. A large number of superpixels would make a superpixelized image too similar to the original image. We aim to model a relatively continuous change from each original image sample to its superpixelized images, from fine to coarse, so that the VRM distribution (or neighborhood distribution) around the original image sample can be better captured. As a result, we choose a range  $[s_l, s_u]$  of values for  $s$ , and form a set of superpixelized images  $SP(x_i, s), s = s_l, \dots, s_u$ , for each original image  $x_i$ .

For biomedical image datasets, the imaging settings are usually known in practice. In particular, the scales and size of the images, and the range of sizes of objects in the images are often known. Thus, one can set the values of  $s_l$  and  $s_u$  based on prior knowledge of these image aspects. For different image sets and applications, one can set  $s_l$  and  $s_u$  differently. In our experiments, for simplicity and for demonstrating the robustness of SPDA, we choose a common setting of  $s_l$  and  $s_u$  for all the 2D segmentation datasets ( $s_l = 800$  and  $s_u = 2000$ ). For the 3D image dataset, due to the increase of image dimensionality, we set  $s_l = 2000$  and  $s_u = 4000$ .

## 2.5. Model training using SPDA

The loss function for training a deep learning based segmentation network using both the original training data and the augmented data is:

$$\frac{1}{n} \sum_{i=1}^n (\mathcal{L}(f(x_i), y_i) + \lambda \sum_{s=s_l}^{s_u} \mathcal{L}(f(SP(x_i, s)), y_i)) \quad (5)$$

where  $\mathcal{L}$  is a spatial cross-entropy loss,  $f$  is the segmentation model under training,  $SP$  is for generating a superpixelized image, and  $s$  is a parameter for  $SP$  that specifies how many superpixels are desired to be generated. We set  $\lambda$  as simple as a normalization term  $\frac{1}{s_u - s_l + 1}$ . We aim to minimize the above function with respect to the parameters of  $f$ .

A common way of optimizing the objective function above is to use a mini-batch based stochastic gradient descent method. Following the loss function in Eq. (5), half of the total samples in the

mini-batch is drawn from the original image samples and the other half is from the SPDA-generated samples. We provide the pseudo-code (**Algorithm 1**) for the model training procedure below.

---

**Algorithm 1:** Model training using SPDA-augmented training data

---

**Data:**  $(x_i, y_i)$  and  $(SP(x_i, s), y_i)$ ,  $i = 1, 2, \dots, n$  and  $s = s_l, \dots, s_u$ .

**Result:** A trained FCN model.

Initialize an FCN model with random weights, mini-batch =  $\emptyset$ ;

**while** stopping condition not met **do**

**for**  $m = 1$  to batch-size/2 **do**

$p = \text{random.randint}(1, n)$ ;

add  $(x_p, y_p)$  to the mini-batch;

$k = \text{random.randint}(s_l, s_u)$ ;

add  $(SP(x_p, k), y_p)$  to the mini-batch;

**end**

Update FCN using data in the mini-batch using the Adam optimizer;

mini-batch =  $\emptyset$ ;

**end**

---

### 3. Experiments

Four biomedical image segmentation datasets are used to evaluate our SPDA method. These datasets are: (1) 3D magnetic resonance (MR) images of myocardium and great vessels (blood pool) in cardiovascular (Pace et al., 2015), (2) electron micrographs (EM) of neuronal structures (Lee et al., 2015), (3) an in-house 2D electron micrographs (EM) of fungal cells that invade animal (ant) tissues, and (4) 2D H&E stained histology images of glands (Sirinukunwattana et al., 2017). Note that SPDA can be extended to segmentation of 3D images using a straightforward extension of SLIC (Achanta et al., 2012) that generates supervoxels instead of superpixels.

On the 2D segmentation datasets, our experiments of SPDA use two common FCN models: U-Net (Ronneberger et al., 2015) and DCN (Chen et al., 2016a). In addition to showing the effectiveness of SPDA, on the neuronal structure and fungus datasets, we also compare SPDA with the elastic deformation for data augmentation (EDDA) used in (Ronneberger et al., 2015). On the 3D segmentation dataset, a state-of-the-art DenseVoxNet (Yu et al., 2017) is utilized for experiments with our SPDA. Experiments on this 3D dataset aim to show the capability of SPDA for 3D image data.

We made a simple extension of the original DCN model (Chen et al., 2016a), which now contains 5 max-pooling layers (deeper than the original DCN) (see Fig. 3). The extension allows DCN to have a larger receptive field for making use of higher-level image information. Random cropping, flipping, and rotation are applied as standard/basic data augmentation operations to all the instances in the experiments. We denote this set of basic data augmentation operations as  $DA_{basic}$ . For fair comparison, we keep all the training settings (e.g., random seed, learning rate, mini-batch size, etc) the same for all the model training. Adam (Kingma and Ba, 2014) optimizer is used for model optimization. As in a common practice, the learning rate for model training is set as 0.0005 for the first 30000 iterations, and then decays to 0.00005 for the rest of the training. The mini-batch size is set as 8. The compactness parameter for SLIC (Achanta et al., 2012) is set as its default value 20. The

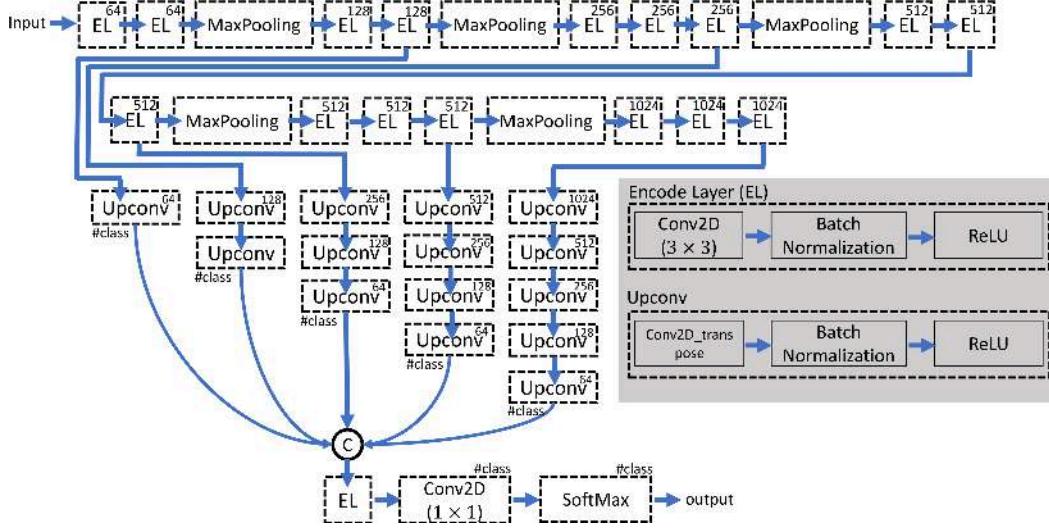


Figure 3: The architecture of the deeper DCN model (which we modify from the original DCN (Chen et al., 2016a)).

size of the input and output of an FCN model is set as  $192 \times 192$  for 2D images and  $64 \times 64 \times 64$  for 3D images. The training procedure stops its execution when there is no significant change in the training errors.

**3D cardiovascular segmentation.** The HVSMR dataset (Pace et al., 2015) was used for segmenting myocardium and great vessels (blood pool) in 3D cardiovascular magnetic resonance (MR) images. The original training dataset contains 10 3D MR images, and the test data consist of another 10 3D MR images. The ground truth of the test data is not available to the public; the evaluations are done by submitting segmentation results to the organizers' server.

The Dice coefficient, average distance of boundaries (ADB), and symmetric Hausdroff distance are the criteria for evaluating the quality of the segmentation results. A combined score  $S$ , computed as  $S = \sum_{\text{class}} (\frac{1}{2} \text{Dice} - \frac{1}{4} \text{ADB} - \frac{1}{30} \text{Hausdorff})$ , is used by the organizers, and this score aims to measure the overall quality of the segmentation results.

When applying SPDA to 3D image data, supervoxels (instead of superpixels) are generated. We use a 3D version of SLIC for generating supervoxels. SPDA is tested using the DenseVoxNet (Yu et al., 2017), which is a state-of-the-art FCN for 3D voxel segmentation. In Table 1, we show the results from DenseVoxNet +  $DA_{\text{basic}}$ , DenseVoxNet +  $DA_{\text{basic}}$  + SPDA, and other known models on this dataset. One can see that SPDA improves the segmentation results significantly, especially on the average distance of boundaries and Hausdroff distance metrics.

**Fungal segmentation.** We further evaluate SPDA using an in-house EM fungus dataset that contains 6 large 2D EM images ( $4000 \times 4000$  each) for segmentation. Since the input window size of a fully convolutional network is set as  $192 \times 192$ , there are virtually hundreds and thousands unique image samples for model training and testing. This dataset contains three classes of objects of interest: fungal cells, muscles, and nervous tissue. We use 1 large microscopy image for training, and 5 large microscopy images for testing. This experiment aims to evaluate the effectiveness of SPDA in a difficult situation in which the training set is smaller than the test set (not uncommon

Table 1: Comparison of segmentation results on the HVSMR dataset.

Method	Myocardium			Blood pool			Overall score
	Dice	ADB	Hausdorff	Dice	ADB	Hausdorff	
3D U-Net ( <a href="#">Çiçek et al., 2016</a> )	0.694	1.461	10.221	0.926	0.940	8.628	-0.419
VoxResNet ( <a href="#">Chen et al., 2018</a> )	0.774	1.026	6.572	0.929	0.981	9.966	-0.202
DenseVoxNet ( <a href="#">Yu et al., 2017</a> ) + $DA_{basic}$	<b>0.821</b>	0.964	7.294	0.931	0.938	9.533	-0.161
DenseVoxNet ( <a href="#">Yu et al., 2017</a> ) + $DA_{basic}$ + SPDA	0.817	<b>0.723</b>	<b>3.639</b>	<b>0.938</b>	<b>0.778</b>	<b>5.548</b>	<b>0.196</b>

in biomedical image segmentation). In Table 2, Student’s t-test suggests that all our improvements are significant. The p-values for MeanIU of U-Net vs U-Net + SPDA, U-Net + EDDA vs U-Net + SPDA, DCN vs DCN + SPDA, and DCN + EDDA vs DCN + SPDA are all  $< 0.0001$ .

Table 2: Comparison results on the fungus segmentation dataset: The Intersection-over-Union (IoU) scores for each object class and the MeanIU scores across all the classes of objects. U-Net ([Ronneberger et al., 2015](#)), DCN ([Chen et al., 2016a](#)), and EDDA: elastic deformation data augmentation (used in ([Ronneberger et al., 2015](#))) are considered.

Method	Fungus	Muscle	Nervous tissue	MeanIU
U-Net + $DA_{basic}$	$0.849 \pm 0.008$	$0.976 \pm 0.003$	$0.506 \pm 0.029$	$0.777 \pm 0.008$
U-Net + $DA_{basic}$ + EDDA	$0.881 \pm 0.007$	$0.975 \pm 0.004$	$0.549 \pm 0.035$	$0.8019 \pm 0.014$
U-Net + $DA_{basic}$ + SPDA	$0.927 \pm 0.001$	$0.973 \pm 0.002$	$0.667 \pm 0.020$	<b><math>0.856 \pm 0.007</math></b>
DCN + $DA_{basic}$	$0.783 \pm 0.064$	$0.970 \pm 0.009$	$0.349 \pm 0.092$	$0.701 \pm 0.055$
DCN + $DA_{basic}$ + EDDA	$0.863 \pm 0.042$	$0.970 \pm 0.008$	$0.453 \pm 0.183$	$0.762 \pm 0.078$
DCN + $DA_{basic}$ + SPDA	$0.907 \pm 0.011$	$0.973 \pm 0.005$	$0.630 \pm 0.026$	<b><math>0.837 \pm 0.012</math></b>

**Neuronal structure segmentation.** We experiment with SPDA using the EM mouse brain neuronal images ([Lee et al., 2015](#)). This dataset contains 4 stacks of EM images (1st:  $255 \times 255 \times 168$ , 2nd:  $512 \times 512 \times 170$ , 3rd:  $512 \times 512 \times 169$ , and 4th:  $256 \times 256 \times 121$ ). Following the practice in ([Lee et al., 2015](#); [Shen et al., 2017](#)), we use the 2nd, 3rd, and 4th stacks for model training and the 1st stack for testing. Since the image stacks in this dataset are highly anisotropic (i.e., the voxel spacing along the  $z$ -axis is much larger than those along the  $x$ - and  $y$ -axes), directly applying 3D models with 3D convolutions is not very suitable for highly anisotropic 3D images. Hence, for simplicity, our experiments on this dataset are based on superpixels in the 2D slices of the 3D images and using 2D FCN models, instead of supervoxels and 3D models. We run all experiments 5 times with different random seeds. The average performance across all the runs and their standard deviations are reported in Table 3. Student’s t-test suggests that all our improvements are significant. The p-values for  $V_{Fscore}^{Rand}$  are:  $< 0.0001$  for U-Net vs U-Net + SPDA, 0.0059 for U-Net + EDDA vs U-Net + SPDA,  $< 0.0001$  for DCN vs DCN + SPDA, and 0.0042 for DCN + EDDA vs DCN + SPDA.

**Gland segmentation.** This H&E stained microscopy image dataset ([Sirinukunwattana et al., 2017](#)) contains 85 training images (37 benign (BN), 48 malignant (MT)) and 60 testing images (33 BN, 27 MT) in part A, and 20 testing images (4 BN, 16 MT) in part B. Table 4 shows the gland segmentation results that demonstrate the effect of SPDA and comparison with the state-of-the-art

Table 3: Comparison results on the neuronal structure segmentation dataset:  $V^{Rand}$  scores for evaluating the segmentation quality.  $DA_{basic}$ : basic data augmentation operations (random cropping, flipping, and rotation); EDDA: elastic deformation data augmentation in (Ronneberger et al., 2015).

Method	$V_{merge}^{Rand}$	$V_{split}^{Rand}$	$V_{Fscore}^{Rand}$
$M^2$ FCN (Shen et al., 2017)	0.9917	0.9815	0.9866
U-Net (Ronneberger et al., 2015) + $DA_{basic}$	$0.9954 \pm 0.0003$	$0.9879 \pm 0.0001$	$0.9917 \pm 0.0001$
U-Net + $DA_{basic}$ + EDDA	$0.9957 \pm 0.0005$	$0.9931 \pm 0.0003$	$0.9944 \pm 0.003$
U-Net + $DA_{basic}$ + SPDA	<b><math>0.9965 \pm 0.0003</math></b>	<b><math>0.9935 \pm 0.0004</math></b>	<b><math>0.9950 \pm 0.0002</math></b>
DCN (Chen et al., 2016a) + $DA_{basic}$	$0.9950 \pm 0.0003$	$0.9916 \pm 0.0001$	$0.9933 \pm 0.0001$
DCN + $DA_{basic}$ + EDDA	$0.9980 \pm 0.0006$	$0.9917 \pm 0.0001$	$0.9949 \pm 0.0002$
DCN + $DA_{basic}$ + SPDA	<b><math>0.9987 \pm 0.0010</math></b>	<b><math>0.9921 \pm 0.0006</math></b>	<b><math>0.9954 \pm 0.0002</math></b>

Table 4: Comparison results on the gland segmentation dataset: The  $F_1$  score and ObjectDice evaluate how well glands are segmented at the instance level, and Object Hausdorff distance evaluates the shape similarity between the segmented objects and ground truth objects.

Method	$F_1$ Score		ObjectDice		ObjectHausdorff	
	part A	part B	part A	part B	part A	part B
CUMedVision (Chen et al., 2016b)	0.912	0.716	0.897	0.718	45.418	160.347
Multichannel1 (Xu et al., 2016b)	0.858	0.771	0.888	0.815	54.202	129.930
Multichannel2 (Xu et al., 2016a)	0.893	0.843	0.908	0.833	44.129	116.821
MILD-Net (Graham et al., 2018)	0.914	0.844	<b>0.913</b>	0.836	<b>41.54</b>	105.89
U-Net (Ronneberger et al., 2015) + $DA_{basic}$	0.89202	0.8087	0.88193	0.83441	51.19	108.25
U-Net + $DA_{basic}$ + SPDA	0.9007	0.83843	0.88429	0.8415	49.95	107.69
DCN (Chen et al., 2016a) + $DA_{basic}$	0.9071	0.825	0.898	0.826	48.740	126.479
DCN + $DA_{basic}$ + SPDA	<b>0.918</b>	<b>0.860</b>	<b>0.913</b>	<b>0.858</b>	42.620	<b>95.83</b>

models on this dataset. In particular, using SPDA, DCN can be trained to perform considerably better than the state-of-the-art model (Graham et al., 2018).

## 4. Conclusions

In this paper, we presented a new data augmentation method using superpixels (or supervoxels), SPDA, for training fully convolutional networks for biomedical image segmentation. Our proposed SPDA method is well motivated, easy to use, compatible with known data augmentation techniques, and can effectively improve the performance of deep learning models for biomedical image segmentation tasks.

## References

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transac-*

- tions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.
- Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.
- Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In *Advances in Neural Information Processing Systems*, pages 416–422, 2001.
- Hao Chen, Xiaojuan Qi, Jie-Zhi Cheng, Pheng-Ann Heng, et al. Deep contextual networks for neuronal structure segmentation. In *AAAI*, pages 1167–1173, 2016a.
- Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. DCAN: Deep contour-aware networks for accurate gland segmentation. In *CVPR*, pages 2487–2496, 2016b.
- Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, 170:446–455, 2018.
- Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *MICCAI*, pages 424–432, 2016.
- Zach Eaton-Rosen, Felix Bragman, Sebastien Ourselin, and M Jorge Cardoso. Improving data augmentation for medical image segmentation. In *MDL*, 2018.
- Simon Graham, Hao Chen, Qi Dou, Pheng-Ann Heng, and Nasir Rajpoot. MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *arXiv preprint arXiv:1806.01963*, 2018.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
- Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- Kisuk Lee, Aleksandar Zlateski, Vishwanathan Ashwin, and H Sebastian Seung. Recursive training of 2D-3D convolutional networks for neuronal boundary prediction. In *NIPS*, pages 3573–3581, 2015.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- Danielle F Pace, Adrian V Dalca, Tal Geva, Andrew J Powell, Mehdi H Moghari, and Polina Golland. Interactive whole-heart segmentation in congenital heart disease. In *MICCAI*, pages 80–88, 2015.

- Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.
- Wei Shen, Bin Wang, Yuan Jiang, Yan Wang, and Alan Yuille. Multi-stage multi-recursive-input fully convolutional networks for neuronal boundary detection. In *ICCV*, pages 2410–2419, 2017.
- Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, et al. Gland segmentation in colon histology images: The GlaS challenge contest. *Medical Image Analysis*, 35:489–502, 2017.
- Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. *arXiv preprint arXiv:1801.05401*, 2018.
- Yan Xu, Yang Li, Mingyuan Liu, Yipei Wang, Yubo Fan, Maode Lai, Eric I Chang, et al. Gland instance segmentation by deep multichannel neural networks. *arXiv preprint arXiv:1607.04889*, 2016a.
- Yan Xu, Yang Li, Mingyuan Liu, Yipei Wang, Maode Lai, I Eric, and Chao Chang. Gland instance segmentation by deep multichannel side supervision. In *MICCAI*, pages 496–504, 2016b.
- Lequan Yu, Jie-Zhi Cheng, Qi Dou, Xin Yang, Hao Chen, Jing Qin, and Pheng-Ann Heng. Automatic 3D cardiovascular MR segmentation with densely-connected volumetric ConvNets. In *MICCAI*, pages 287–295, 2017.
- Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

## Appendix A. Empirical Studies of SPDA-generated Data

In this appendix, two empirical studies are conducted to show: (1) the SPDA-generated data are “near” their original image data in the data space, and (2) the data distribution of the SPDA-augmented training set is “closer” to the distribution of the test data (or true data).

### A.1. SPDA-generated data near their original samples

We seek to examine how the SPDA-generated data are spatially close to their corresponding original images. Since  $SP(x_i, s)$  and  $x_i$  are both in a high dimensional space, comparing them is not a trivial task. One may use a distance metric for measuring the distance between  $SP(x_i, s)$  and  $x_i$ . However, with different metrics, the meaning of “being different” or “being similar” can be drastically different.

To avoid too much complication in manifold learning or metric learning, we use two common dimensionality reduction methods, standard PCA and t-SNE (Maaten and Hinton, 2008), to help visualize the original image samples and SPDA-generated image samples. Fig. 4 shows visualization results of such samples (both the original and SPDA-generated samples) on the neuronal structure dataset, fungus dataset, and gland dataset (after applying PCA). One may observe that the SPDA-generated data are near/surrounding the original image data, forming a close neighborhood of the original images. In Fig. 5, we provide visualization views of some SPDA-generated image samples using t-SNE (Maaten and Hinton, 2008).

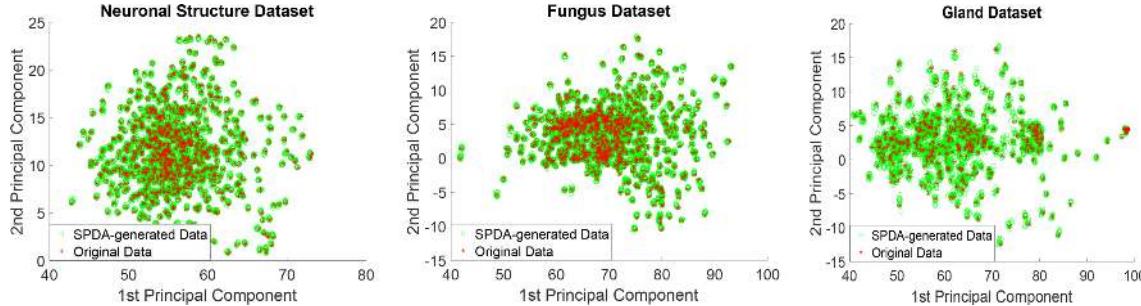


Figure 4: After dimensionality reduction using PCA, each original image sample (red) is surrounded by (or closely adjacent to) its corresponding superpixelized images (green). Zoom-in view would show more details.

### A.2. Data distribution comparison

Here we are interested in a basic question: Whether adding SPDA-generated data  $X_{spda}$  to the original training set  $X_{ori}$  makes the new training set  $X_{augmented}$  “closer” to the test data  $X_{test}$  in the image representation space.

We utilize variational auto-encoders (VAEs) (Kingma and Welling, 2013) to encode the training images  $X = \{x_1, \dots, x_n\}$  into much lower dimensional representation  $Z = \{z_1, \dots, z_n\}$ . On each dimension of the space thus resulted, the data are expected to follow a Gaussian distribution with zero mean and unit variance. This is a standard objective of VAE.

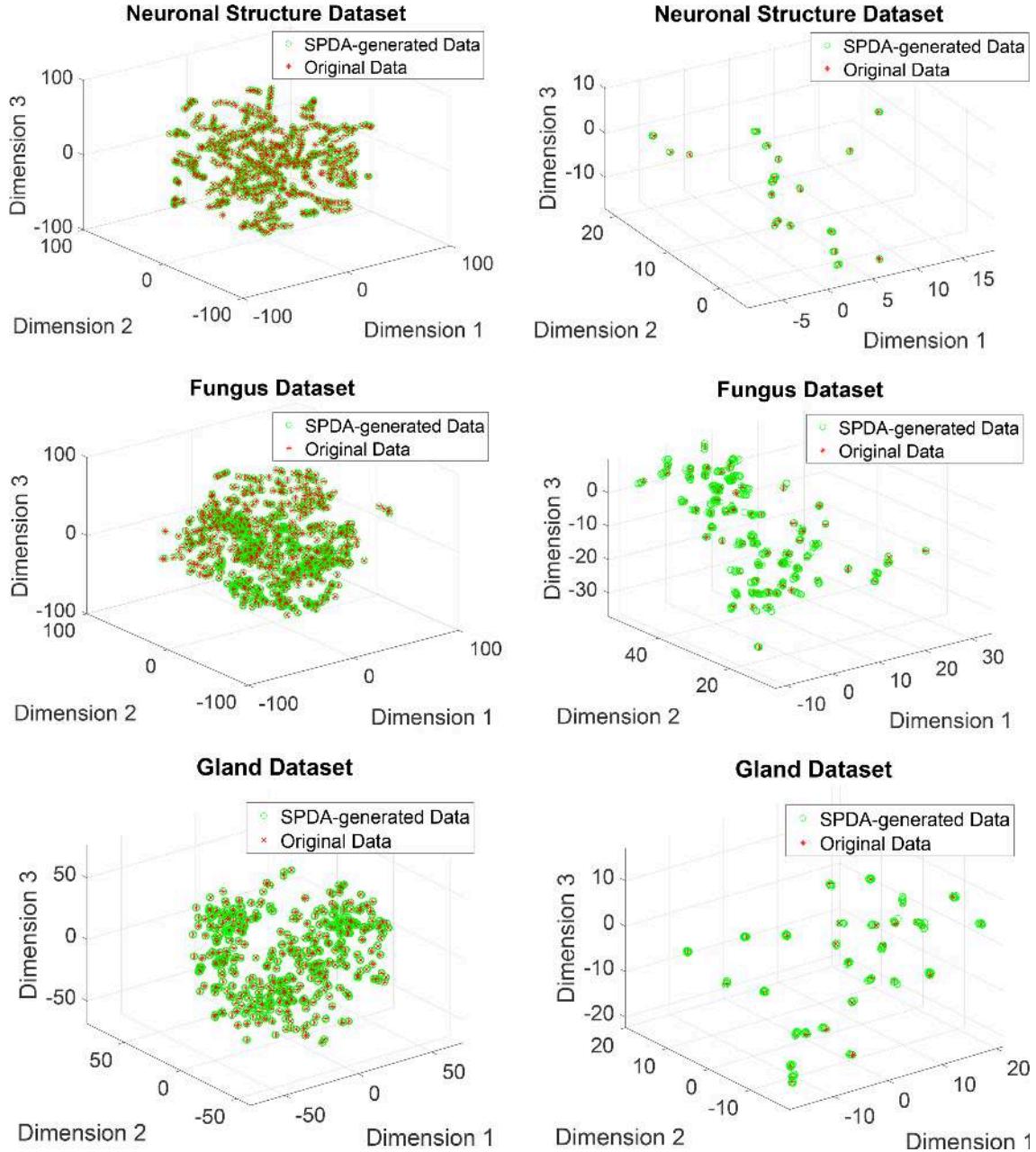


Figure 5: Visualization of some SPDA-generated image samples (green) and the original training samples (red) using t-SNE (Maaten and Hinton, 2008). Left: Overview of the samples; right: zoom-in views. SPDA-generated samples are in a close neighborhood of their corresponding original samples.

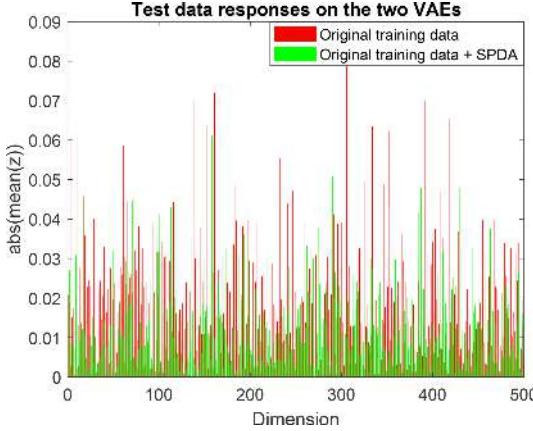


Figure 6: The responses of the test data on the two VAEs trained using the original training data and the SPDA-augmented training data (on the neuronal structure dataset).

To show the effect of SPDA, we train two VAEs: VAE-A is trained using only the original training images  $X_{ori}$ , and VAE-B is trained using the SPDA-augmented training set  $X_{augmented} = X_{ori} \cup X_{spda}$ . These two VAEs are all trained using the same settings; the only difference is their training data. After training, VAE-A is applied to its training data  $X_{ori}$  and the test data  $X_{test}$ , to obtain  $Z_{ori}^A$  and  $Z_{test}^A$ . Similarly, VAE-B is applied to its training data  $X_{augmented}$  and the test data  $X_{test}$ , and  $Z_{augmented}^B$  and  $Z_{test}^B$  are obtained. We then compare

$$D_{KL}(P(z_{test}^A) || P(z_{ori}^A)) \quad (6)$$

with

$$D_{KL}(P(z_{test}^B) || P(z_{augmented}^B)) \quad (7)$$

and compare

$$D_{KL}(P(z_{ori}^A) || P(z_{test}^A)) \quad (8)$$

with

$$D_{KL}(P(z_{augmented}^B) || P(z_{test}^B)) \quad (9)$$

where  $D_{KL}$  is the Kullback-Leibler divergence (Kullback and Leibler, 1951) and  $P(z)$  is the probability distribution of  $z$ . The above procedure is applied to the neuronal structure dataset. The results are:

$$D_{KL}(P(z_{test}^B) || P(z_{augmented}^B)) = 5.5517 < D_{KL}(P(z_{test}^A) || P(z_{ori}^A)) = 5.8779,$$

and

$$D_{KL}(P(z_{augmented}^B) || P(z_{test}^B)) = 5.0586 < D_{KL}(P(z_{ori}^A) || P(z_{test}^A)) = 6.1491.$$

It is clear that SPDA can potentially make the training data distribution closer to the test data/true data distribution in the image representation space. We believe this is a main reason why learning models trained using SPDA-augmented training data can generalize better on test data. To show this observation visually, the absolute values of the averages of  $Z_{test}^A$  and  $Z_{test}^B$  are shown in Fig 6. One

can see that the values of  $Z_{test}^B$  are generally closer to 0 than  $Z_{test}^A$ , which means that the distribution of  $Z_{test}^B$  is closer to the zero mean Gaussian distribution, and thus is closer to the distribution of  $Z_{augmented}^B$ .

# CARE: Class Attention to Regions of Lesion for Classification on Imbalanced Data

Jiaxin Zhuang<sup>\*1</sup>

Jiabin Cai<sup>\*1</sup>

Ruixuan Wang<sup>1</sup>

Jianguo Zhang<sup>2</sup>

Weishi Zheng<sup>1</sup>

ZHUANGJX5@MAIL2.SYSU.EDU.CN

CAIJB5@MAIL2.SYSU.EDU.CN

WANGRUIX5@MAIL.SYSU.EDU.CN

J.N.ZHANG@DUNDEE.AC.UK

WSZHENG@IEEE.ORG

<sup>1</sup> School of Data and Computer Science , Sun Yat-sen University, China

<sup>2</sup> Computing, School of Science and Engineering, University of Dundee, UK

## Abstract

To date, it is still an open and challenging problem for intelligent diagnosis systems to effectively learn from imbalanced data, especially with large samples of common diseases and much smaller samples of rare ones. Inspired by the process of human learning, this paper proposes a novel and effective way to embed attention into the machine learning process, particularly for learning characteristics of rare diseases. This approach does not change architectures of the original CNN classifiers and therefore can directly plug and play for any existing CNN architecture. Comprehensive experiments on a skin lesion dataset and a pneumonia chest X-ray dataset showed that paying attention to lesion regions of rare diseases during learning not only improved the classification performance on rare diseases, but also on the mean class accuracy.

**Keywords:** Attention, Imbalanced Data, Small Samples, Skin Lesion, Pneumonia Chest X-ray.

## 1. Introduction

Deep learning has been widely applied to computer-aided diagnosis systems, particularly based on medical images (Shen et al., 2017). However, human-level performance of intelligent diagnosis often comes from training deep neural networks on large automated data. Therefore, current intelligent systems are mainly trained for the diagnosis of commonly encountered diseases. To date, due to the limited available data for rare diseases, it is still an open and challenging problem to train an intelligent system for diagnosis of both common and rare diseases. To solve this problem, the key is how to effectively handle the data imbalance between common diseases and rare ones.

Multiple approaches have been proposed to solve such data imbalance problems. One traditional approach is through over-sampling of the limited data for small-sample classes or down-sampling of the data for larger-sample classes(Chawla et al., 2002), thus generating similar number of training data between classes. Data augmentation, a default choice for training deep neural networks, can also be used as an over-sampling method to generate more data for small-sample classes. Another widely used approach is to improve the cost of mis-classifying each training example coming from small-sample classes, which can be easily realized by setting larger weights for small-sample classes in the loss function (Sun et al., 2007). Different from setting a single class weight for all training examples of the same class, another type of approach is to adaptively re-weight each single training

---

\* Contributed equally

example based on the difficulty of being correctly classified, including boosting (Dollár et al., 2009; Freund and Schapire, 1999; Viola and Jones, 2001) and the recently proposed focal loss (Lin et al., 2017). Based on such weights, hard negative mining can be adopted to select just a subset of training data for the next-round training of classifier (Dalal and Triggs, 2005; Felzenszwalb et al., 2010; Fu et al., 2017; Shrivastava et al., 2016; Sung, 1995). Besides these approaches, particularly for medical image analysis, transfer learning via fine-tuning a pre-trained classifier has been proven helpful to improve performance for both large- and small-sample classes (Wang and Xia, 2018; Buda et al., 2017).

Different from these existing studies which consider each image as the basic unit and mainly focus on varying number and importance of images, this paper proposes a novel approach to data imbalance problems by delving into images and considering the high-level semantics of images, i.e, Class Attention to REgions of lesion (we termed our approach as CARE). Specifically, inspired by the process of human learning, attention was embedded into the learning process of neural network classifiers particularly for rare diseases. By attracting classifiers to pay more attention to the lesion regions during learning, the classifiers can learn more effectively from small samples. Due to limited training data for rare diseases, annotation of lesion regions (in the form of bounding boxes containing lesion regions) does not usually take much effort for radiologists and therefore is reasonably acceptable. Different from existing attention-relevant deep learning studies where attention is estimated as intermediate outputs of neural networks (Vaswani et al., 2017; Xu et al., 2015), the proposed approach provides a novel way to explicitly uses attention as part of supervision signal (in addition to image labels) to help train classifiers. What's more, the proposed attention embedding mechanism is independent of and does not alter neural network architectures, therefore can directly used as an element for any existing convolutional neural network architecture. In addition, the CARE approach is independent of any existing approach to data imbalance, and therefore can be combined to handle the imbalance problem together. Experiments with multiple different neural network architectures on a skin lesion dataset and a pneumonia chest X-ray dataset showed that paying attention to lesion regions of rare diseases during learning did improve the classification performance on rare diseases. Compared to existing approaches, addition of the CARE approach always further improved performance.

## 2. The CARE Approach

Medical students are often pointed at the regions lesions containing distinct characteristics of certain diseases when taught to diagnose diseases via medical images (Krupinski, 2010). With the help of such attentions to lesion regions, students probably can more effectively learn to grasp the distinct properties of each disease even with a small sample of medical images. Inspired by the learning process of humans, here we propose a simple but effective method to embed attention into the learning process of deep neural network classifiers for intelligent diagnosis.

We hypothesize that appropriate attention during learning would help neural network classifiers more effectively learn from small samples particularly for rare diseases. Suppose the lesion regions of interest have been provided in advance for model learning, in the form of bounding boxes containing lesions. The effort of providing bounding boxes is feasible for rare diseases because quite often only a small sample of images are available for each category. Then, if there is one way to estimate the local regions on which the classifier focuses during image diagnosis, by enforcing that such ‘visual focus’ of the classifier falls into the bounded lesion regions, attention would be

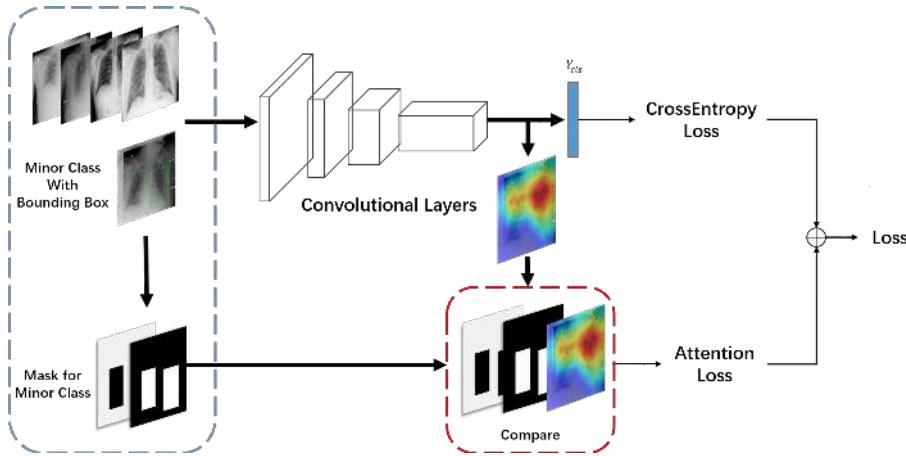


Figure 1: The diagram of the proposed CARE framework with the attention loss. It contains two branches with one focusing on attention into lesion region in minority class. Bounding boxes are provided only for minority classes during training.

naturally embedded during learning. Fortunately, such ‘visual focus’ of a classifier on any input image can be conveniently estimated by a recently proposed visualization approach called Grad-CAM (Selvaraju et al., 2016). Given a well trained classifier and an input image, Grad-CAM can provide a class-specific feature activation map in which regions with higher activation contribute more to the classifier’s output prediction being the specific class. Therefore, if the classifier attends to only the box-bounded regions when diagnosing an image, the high activation regions from the Grad-CAM should also be within the bounded regions. In this sense, the spatial relationship (e.g., degree of overlap) between the high activation regions and the bounded image regions can be used to measure how well the classifier has attended to the bounded image regions.

Denote by  $L_a$  the discrepancy between the high activation regions from Grad-CAM and the bounded image regions over all training data, then embedding attention during classifier learning can be realized by minimizing a new loss  $L$  for the classifier,

$$L = (1 - \alpha)L_c + \alpha L_a \quad (1)$$

where  $L_c$  is the general cross-entropy loss for the network classifier, and  $L_a$ , called *attention loss*, helps to drive the network to attend to box-bounded image regions during training.  $\alpha$  is a coefficient to balance the two loss terms. Considering the different influence of the inside-box and outside-box regions, the attention loss is further split into two items by

$$L_a = L_{in} + \lambda L_{out} \quad (2)$$

where the inner loss  $L_{in}$  helps the classifier increase the attention inside the bounding box, and the outer loss  $L_{out}$  helps the classifier decrease the attention outside the bounding box.  $\lambda$  is a coefficient to balance the two loss terms. In detail, for any training image with bounding box(es) provided, let  $M_{in}$  denote a binary complement image in which all pixels inside bounding box are set to 1 and others to 0, and in contrast,  $M_{out}$  denote a binary mask image in which all pixels inside bounding box are set to 0 and any pixel outside box is set to either 1 or a positive value relevant to the distance

between the pixel and the bounding box. Let  $F$  denote the feature activation map from Grad-CAM for the training image based on current classifier. Then  $L_{in}$  and  $L_{out}$  (for one training image) can be defined as

$$L_{in} = -\min\left(\frac{\sum_{i,j} M_{in}(i,j) \cdot F(i,j)}{\sum_{i,j} M_{in}(i,j)}, \tau\right) \quad (3)$$

$$L_{out} = \frac{\sum_{i,j} M_{out}(i,j) \cdot F(i,j)}{\sum_{i,j} M_{out}(i,j)} \quad (4)$$

Here,  $M_{in}(i,j)$  represents the value at the position  $(i,j)$  in the mask  $M_{in}$ , and similarly for  $M_{out}(i,j)$  and  $F(i,j)$ . Equation (4) represents the strength of feature activation outside the bounding box, while Equation (3) would penalize the classifier if the highly activated area inside the bounding box is not large enough (i.e., when the percent of weighted activated area  $\frac{\sum_{i,j} M_{in}(i,j) \cdot F(i,j)}{\sum_{i,j} M_{in}(i,j)}$  is smaller than a predefined threshold  $\tau$ ). Note that for notation simplicity, Equations (3) and (4) are just for one single image. In fact, during training, the loss terms are calculated and averaged over all training images.

One advantage of the proposed attention-based approach is its independence of model structures. Therefore the CARE can be directly embedded to the training processing of any existing CNN classifiers, without alternating their model architectures. Also, the CARE framework is independent of existing approaches to handling data imbalance, therefore can be directly combined to further improve classification performance.

### 3. Experiment

#### 3.1. Experimental settings

**Dataset.** Two medical image datasets were used to evaluate the proposed approach. One is the skin dataset provided by the ISIC2018 Challenge with 7 disease categories (Codella et al., 2017), in which 6705 images are for Melanocytic nevus and only 115 images for Dermatofibroma, clearly having serious data imbalance between classes. One bounding box was generated for each image of the rare disease Dermatofibroma by one of the authors and confirmed by a dermatologist. The other is the pneumonia detection X-ray dataset with 3 categories <sup>1</sup>, including 8,851 ‘Normal’ images, 6012 ‘Lung Opacity’ images, and images of ‘No Lung Opacity/Not Normal’. Each ‘Lung Opacity’ image was provided with one or multiple bounding boxes indicating the region of the pneumonia. Although the original objective of this chest X-ray data is for lesion detection, we used it for 3-class classification, with the ground-truth bounding boxes used to evaluate the proposed approach. The number of ‘Lung Opacity’ images is much smaller than other two categories, being considered as the minority class in a data imbalance scenario. All images were resized to  $224 \times 224$  pixels, with bounding boxes resized accordingly for the small-sample class in each dataset. For each dataset, images are randomly split into training set (80%) and test set (20%) with stratification.

**Implementation and Protocol.** In the experiments, the training of CARE is divided into two stages. At the first stage, each backbone CNN classifier (i.e., the branch without the attention loss in Figure 1) used was pretrained first on ImageNet and then on the training set without the attention loss. The training at this stage is stopped when the cross-entropy loss does not decrease any more (normally within 200 epochs in our experiment). At the second stage, the attention loss was included

---

1. The original dataset comes from <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge>, and part of the dataset was extracted for the purpose of evaluation on data imbalance.

to fine-tune the stage-one classifier with the training set.  $\alpha$  in Equation(1) was set to 0.5 unless otherwise mentioned. Adam optimizer was used throughout, with initial learning rate set at 0.0001.  $\lambda$  was empirically set to 0.5 for X-ray dataset and 0.7 for Skin Lesion dataset. Specially, we form batches by randomly sampling P classes. In testing, each test image (without any bounding box) was fed to the backbone CNN classifier for prediction. Note that the CARE approach needs no bounding box in testing. *Recall* for the small-sample class (i.e., Dermatofibroma in skin dataset, and Lung Opacity in the pneumonia dataset) and *mean class accuracy* (MCA, i.e., average recall over all classes) were reported as measurement of the model performance.

### 3.2. Results

**Baseline and Comparisons.** In order to test the effectiveness of the proposed approach, we compared CARE to three widely-used strategies for handling data imbalance, namely, 1) *cost sensitive learning* denoted by CSL (Sun et al., 2007), 2) *focal loss* (Lin et al., 2017) denoted by FL, a representative method of hard negative mining, and 3) *data augmentation* (including rotation, flip and color jitter) denoted by DA. We further tested our approach by embedding CARE into the three strategies, resulting methods of CARE+CSL, CARE+FL, and CARE+DA. We also tested a baseline without using the visual attention branch in Figure 1. Table 1 shows the comparison results on the pneumonia and skin datasets. It can be observed that CARE outperforms the baseline significantly in terms of both *recall* and *MCA*, in particular with a large margin on recall for the small-sample class (31.12 vs. 7.41 on the pneumonia dataset, and 52.17 vs. 47.83 on the skin set). All the three strategies (i.e., CSL, FL and DA) perform better than the baseline without any treatment of data imbalance, which is expected. It is worth highlighting that adding CARE to each of CSL, FL or DA could boost the performances significantly w.r.t the use of each method alone; for instance, the recall (and MCA) of CARE+CSL is 45.04 (and 65.23), significantly better than CSL only. For the CSL method, additional experiments showed that varying loss coefficients for the minority class did not change the finding, i.e., CARE+CSL always performs better than CSL alone. This clearly indicates that our approach is capable of boosting their performances significantly when plugged into the existing strategies for handling data imbalance.

**Flexibility with Architecture.** Our proposed CARE framework is independent of model structures. To show this, we test variants of our CARE framework built with two different widely-used CNN architectures: ResNet (He et al., 2015) and VGG19 (Simonyan and Zisserman, 2014). For ResNet, we further test different number of layers at 18, 50, 152, from shallow to very deep. VGG19 uses a 19-layered structure. Thus in total we have four backbones of CNN architectures: ResNet18, ResNet50, ResNet152 and VGG19. For each of the backbones, we compare the performance of the resulting CARE model ( $X(\text{CARE})$ :  $X$  represents the name of the backbone) and the original backbone network; for example VGG19 vs. VGG19+CARE. Results are shown in Table 2. It can be observed that different *original* backbone architectures perform differently, among which VGG-19 performs the best. For each of the backbone, its CARE version outperforms than the original network in terms of both *recall* and *MCA*, with *recall* significantly better.

**Tolerance to Bounding Box.** It is noted that the training of our model needs the bounding box annotations. For many rare or uncommon diseases (such as Dermatofibroma studied in this paper), the annotation effort of bounding boxes (bbox) for the lesion regions in the minority class is usually very small compared to that of accurate boundary pixel-level annotations. Even though, there might exist inter- or intra-observer variations of annotations. The bbox used thus far is tightly around the

Table 1: Comparison on pneumonia dataset and the skin dataset using ResNet50, including baseline, CARE (ours), CSL (cost sensitive learning), FL (focus loss), DA(data augmentation), CARE+CSL (ours), CARE+FL (ours) and CARE+DA(ours). MCA is the mean class accuracy, and recall is reported for the minority class (Lung Opacity/Dermatofibroma).

Pneumonia Dataset			Skin Dataset	
Model	recall(%)	MCA(%)	recall	MCA(%)
baseline	7.41	56.77	47.83	75.75
CARE (ours)	<b>31.12</b>	<b>63.29</b>	<b>52.17</b>	<b>76.16</b>
CSL	11.11	57.88	61.91	80.21
CARE+CSL (ours)	45.04	65.23	<b>65.22</b>	<b>81</b>
FL	11.14	58.41	38.3	72.72
CARE+FL (ours)	<b>49.44</b>	<b>66.72</b>	<b>40.28</b>	<b>74.06</b>
DA	20.06	59.64	56.62	54.41
CARE+DA(ours)	<b>45.18</b>	<b>65.97</b>	<b>60.32</b>	<b>56.22</b>

Table 2: Results of CARE with different backbones on the pneumonia and skin datasets. X(CARE) denotes the CARE model built with backbone X; for instance, VGG19(CARE) represents the CARE model with the backbone model VGG19. Note that all models in the table apply CSL.

Pneumonia Dataset			Skin Dataset	
Model	recall(%)	MCA(%)	recall	MCA(%)
ResNet18	15.16	57.76	58.6	73.71
ResNet18(CARE)	<b>25.51</b>	<b>58.76</b>	<b>59.31</b>	<b>74.06</b>
ResNet50	11.11	57.88	61.91	80.21
ResNet50(CARE)	<b>45.04</b>	<b>65.23</b>	<b>65.22</b>	<b>81</b>
ResNet152	11.37	59.11	61.19	80.15
ResNet152(CARE)	<b>31.31</b>	<b>63.78</b>	<b>72.19</b>	<b>81.93</b>
VGG19	25.93	61.72	52.07	<b>74.48</b>
VGG19(CARE)	<b>41.24</b>	<b>64.3</b>	<b>56.52</b>	72.81

lesion region, which requires the larger annotation effort than other cases of using bbox. To relax this requirement, we vary the bbox by scaling at 0.7, 0.9, 1.0, 1.1 and 1.3, and test the robustness of our approach to such a scaling. Table 3 shows the performance of our model at different scaling. It can be seen that the performance remains stable within a reasonable range, for instance, from 0.9 till 1.3. This indicates that our approach provides certain tolerance to the size of bbox, i.e., the bbox does not need to be tightly around the lesion, with flexibility of using a looser bounding box, which requires less annotation effort.

**Effect of  $\alpha$ .** We further conducted a set of experiments to test the effect of  $\alpha$  using ResNet50(CARE), and the results are shown in Table 4. It could be observed that performances of model remain stable

Table 3: Robustness to BBox scaling, tested with ResNet50(CARE). Note that all models in the table apply CSL.

Pneumonia Dataset			Skin Dataset	
Model	recall(%)	MCA(%)	recall	MCA(%)
without CARE	11.11	57.88	61.91	80.21
0.7	43.56	<b>66.34</b>	63.71	80.20
0.9	<b>50.76</b>	66.27	64.21	80.47
1.0	45.04	65.23	<b>65.22</b>	<b>81</b>
1.1	38.43	63.14	65.22	80.72
1.3	46.06	65.14	65.22	80.68

within a reasonable range, for instance, from 0.1 to 0.9, which indicates that our model is insensitive to the choices of the value of  $\alpha$ . (Jia:)

Table 4: Effect of  $\alpha$  using ResNet50(CARE) on Pneumonia set (left) and Skin set (right). Note that all models in the table apply CSL.

Pneumonia Dataset			Skin Dataset	
Model	recall(%)	MCA(%)	recall	MCA(%)
$\alpha=0$	11.11	57.88	61.91	80.21
$\alpha=0.1$	28.77	63.84	55.22	78.80
$\alpha=0.3$	38.34	64.35	<b>65.22</b>	79.94
$\alpha=0.5$	45.04	<b>65.23</b>	65.21	<b>80.33</b>
$\alpha=0.7$	42.58	64.34	65.18	80.18
$\alpha=0.9$	<b>50.67</b>	65.2	65.12	80.24

**Visual Insight.** To show the effect of the proposed attention loss, we visualize the classification activation maps of the minority class from both datasets, specifically, from the minority class of Lung Opacity, and the minority class of Dermatofibroma. Fig. 2 shows the activation maps of two sample images from each of the classes respectively. For clarity, we also superimpose (ground truth) bounding boxes highlighting the lesion regions on the test images, provided along with the dataset (the Pneumonia dataset) or in-house annotated (the skin dataset). Note that we did *not* use any of those bounding box in the testing, but here merely for visualization purpose. It can be observed that the activated regions (red regions in middle row) without using the proposed attention loss (Eq. 2) clearly deviated from lesion regions, while those (last row) produced by CARE localized the lesion regions well. These results on the test images reveal that the CARE model could have learned to focus on lesion regions when analyzing new images, through attention loss optimized during training.

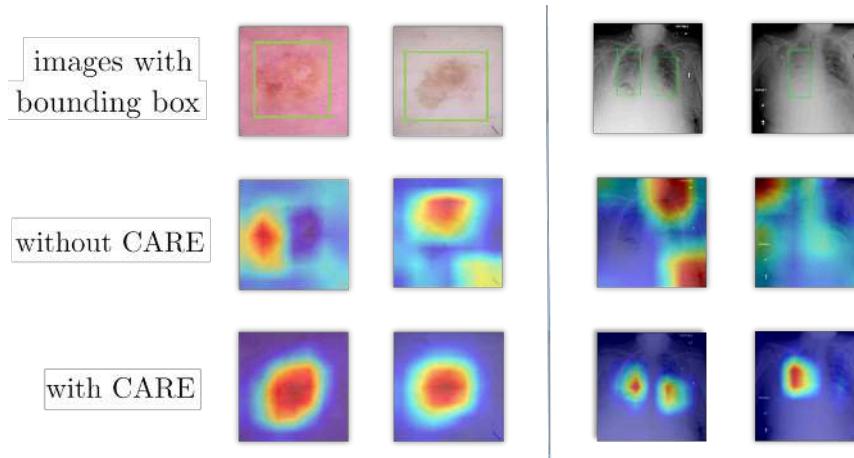


Figure 2: Visualization of activation maps with and without using CARE. Upper row: original images superimposed with bounding boxes; Middle and bottom rows: activation maps without and with using CARE attention loss, respectively. All the class activation maps are generated using Grad-CAM (best viewed in color).

#### 4. Conclusion

We have introduced a novel approach called CARE to embed attention mechanism into CNN learning process, which effectively focuses on minority in the case of data imbalance. This approach, combining Grad-CAM localization and bounding box in minor class indicating the lesion region, is applicable for any CNN based classifier without altering neural network architectures. A series of experiments on the skin and the pneumonia imbalanced datasets have shown our approach can help classifier pay attention to lesion region of rare disease particularly and effectively learn characteristics of diseases from imbalanced data. Our model is effective and can be used to boost the performances of existing strategies of handling data imbalance such as cost sensitive learning, data augmentation, or focal loss.

#### Acknowledgments

This work is supported in part by the National Key Research and Development Plan (grant No. 2018YFC1315402), Royal Society International Exchanges grant (No. 170168), and by the NSFC (grant No. 61628212).

#### References

- Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *CoRR*, 2017.
- Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, pages 321–357, 2002.

- Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging, hosted by the international skin imaging collaboration (ISIC). *CoRR*, 2017.
- Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
- Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge J. Belongie. Integral channel features. In *BMVC*, 2009.
- Pedro F. Felzenszwalb, Ross B. Girshick, and David A. McAllester. Cascade object detection with deformable part models. In *CVPR*, 2010.
- Yoav Freund and Robert E. Schapire. A short introduction to boosting. In *IJCAI*, 1999.
- Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg. DSSD : Deconvolutional single shot detector. *CoRR*, 2017.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, 2015.
- Elizabeth A. Krupinski. Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, 72(5):1205–1217, 2010.
- Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, 2017.
- Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, 2016.
- Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, pages 221–248, 2017.
- Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016.
- K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.
- Yanmin Sun, Mohamed S. Kamel, Andrew K.C. Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, pages 3358 – 3378, 2007.
- Kah Kay Sung. *Learning and example selection for object and pattern detection*. PhD thesis, MIT, 1995.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, 2017.

Paul A. Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

Hongyu Wang and Yong Xia. Chestnet: A deep neural network for classification of thoracic diseases on chest radiography. *CoRR*, 2018.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

