Deep Learning with Limited Labels for Medical Imaging

Mou Cheng Xu

A dissertation submitted in partial fulfillment

of the requirements for

Doctor of Philosophy

of

University College London.

Centre for Medical Image Computing

Department of Medical Physics and Biomedical Engineering

University College London

14th May 2023

I, Mou Cheng Xu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Recent advancements in deep learning-based AI technologies provide an automatic tool to revolutionise medical image computing. Training a deep learning model requires a large amount of labelled data. Acquiring labels for medical images is extremely challenging due to the high cost in terms of both money and time, especially for the pixel-wise segmentation task of volumetric medical scans. However, obtaining unlabelled medical scans is relatively easier compared to acquiring labels for those images.

This work addresses the pervasive issue of limited labels in training deep learning models for medical imaging. It begins by exploring different strategies of entropy regularisation in the joint training of labelled and unlabelled data to reduce the time and cost associated with manual labelling for medical image segmentation. Of particular interest are consistency regularisation and pseudo labelling. Specifically, this work proposes a well-calibrated semi-supervised segmentation framework that utilises consistency regularisation on different morphological feature perturbations, representing a significant step towards safer AI in medical imaging. Furthermore, it reformulates pseudo labelling in semi-supervised learning as an Expectation-Maximisation framework. Building upon this new formulation, the work explains the empirical successes of pseudo labelling and introduces a generalisation of the technique, accompanied by variational inference to learn its true posterior distribution. The applications of pseudo labelling in segmentation tasks are also presented. Lastly, this work explores unsupervised deep learning for parameter estimation of diffusion MRI signals, employing a hierarchical variational clustering framework and representation learning.

Impact Statement

The semi-supervised segmentation methods proposed in this thesis can help to reduce the time required to obtain imaging biomarkers for downstream tasks. For example, in pharmaceutical companies, imaging biomarkers acquired with our methods can enable clinical scientists to more quickly identify the endpoints of drug trials. Similarly, the proposed methods can be used in industries such as self-driving cars and other computer vision-based tasks.

In addition to the direct use of the proposed methods, it is also possible to utilise their components. For instance, the proposed consistency regularisation could be employed to improve the calibration of models for future, safer medical AI systems or other critical real-life applications. The proposed hierarchical variational clustering framework could also be utilised for reconstruction in hyperspectral imaging.

Acknowledgements

First and foremost, I would like to thank my primary supervisor Dr Joseph Jacob. I am forever in his debt for many things, from offering me this PhD, to unconditionally supporting and guiding me in difficult times. I am also extremely appreciative of his open-mindness in letting me explore my own research interests. Thank you Joe, your wisdoms will be my life lessons. I would like to thank my second supervisor, Prof. Daniel Alexander. I need to thank Danny for being such an inspirational leader in the field and for letting me join the CNS group meeting. Thank you Danny for creating this environment. I would also like to thank my third supervisor, Dr Neil Oxtoby. Thank you Neil for being a very good co-author that never bails out at a 2 am deadline. I would never make those conference submissions without your help. I also need to thank my GSK supervisors, Mr. Fred Wilson and Dr. Marius de Groot, for their diligent supports. I apologize to both of you for always bothering you with GSK laptop issues. This PhD wouldn't be fiscally possible without you.

I was as well very fortunate to have two legends in our field as my VIVA examiners, Prof. Julia Schnabel and Prof. Matthew Clarkson. Thank you Julia and Matt, for your meticulous scrutiny of my thesis.

My special thanks goes to my amazing collaborators. I learnt so much from every single one of you, in no particular order: Mr. Yukun Zhou, Dr. Paddy Slator, Dr. Yipeng Hu, Dr. Chen Jin, Dr. Ryutaro Tanno, Dr. Ashkan Pakzad and Miss. An Zhao. Thank you for helping me to complete my half baked ideas and thank you for letting me join your interesting projects. I will never forget about the days of inspiring fruitful discussions and hardcore working before the deadlines. I would also like to thank everyone at CMIC for making the office a fun working environment, and my other friends from old times.

Lastly, I thank my parents in China for raising me and supporting me for my education.

Contents

1 Introduction 18

1.1 Research problem 18

1.2 Structure of the thesis 18

1.3 Publications Covered in this Thesis 19

1.4 Publications Not Covered in this Thesis 20

2 Clinical Background 22

2.1 Imaging Biomarkers in Drug Discovery 22

2.2 Lung and Idiopathic Pulmonary Fibrosis (IPF) 23

2.3 Idiopathic Pulmonary Fibrosis 24

2.4 Treatment and Diagnosis of IPF 26

2.5 Clinical markers for IPF disease progression 26

2.6 Imaging markers for IPF disease progression 27

2.6.1 Emphysema 29

2.6.2 Lung volume 29

2.6.3 Pulmonary vascular structures 30

2.6.4 Traction bronchiecasis 30

2.7 Limitations of the existing quantification tool 30

2.8 Change of the research focus of this PhD 31

3 Machine Learning Background Information 33

3.1 Probability theory 33

3.1.1 Definitions of probability and events 33

3.1.2 Random variables 34

3.1.3 Derivations of Mean and Variance 35

3.1.4 Distributions 38

3.1.5 Bayesian statistics 38

3.1.6 Expectation Maximization 39

3.2 Deep Learning 40

3.2.1 XOR problem 40

3.2.2 Brief introduction of deep learning 42

3.2.3 Layers 43

3.2.4 Architectures 46

3.2.5 Loss functions 48

3.2.6 Backpropagation 48

3.2.7 Stochastic Gradient Descent 50

3.2.8 Training tricks 51

3.3 Deep Learning for Medical Image Segmentation 52

3.3.1 Supervised Learning 53

3.3.2 Semi-Supervised Learning 54

3.3.2.1 SSL in Classification 55

3.3.2.2 SSL in Segmentation 55

3.3.2.3 Mutual Information and entropy minimisation . . 56

4 MisMatch: Calibrated Segmentation via Consistency on Differential Morphological Feature Perturbations with Limited Labels 58

4.1 Abstract 58

4.2 Introduction 59

4.3 Motivations 62

4.4 Methods 65

4.4.1 Background: ERF and the foreground 65

4.4.2 Architecture of Mismatch 67

4.4.3 Positive Attention Shifting Block 67

4.4.3.1 ERF size in Positive Attention Shifting Block . . 68

4.4.4 Negative Attention Shifting Block 68
4.4.4.1 ERF size in Negative Attention Shifting Block . . 68
4.4.5 Loss Functions 69
4.5 Experiments 69
4.5.1 Data sets & Pre-processing 70
4.5.2 Implementation 71
4.5.3 Baselines 71
4.6 Segmentation Results 73
4.6.1 Ablation Studies 75
4.7 Visualisation of the effectiveness of Learnt Attention Masks 77
4.8 Confidence and Calibration of Mismatch 80
4.9 Conclusion 82

5 Expectation-Maximization Pseudo Labelling for Segmentation with Limited Annotations 83

5.1 Abstract 83
5.2 Introduction 84
5.2.1 Label scarcity: one major bottleneck of deep learning in medical imaging 84
5.2.2 Brief review of semi-supervised learning 85
5.2.3 Motivations and contributions 87
5.3 Related works 88
5.3.1 Pseudo Labelling 88
5.3.2 Semi-supervised learning in medical image segmentation . . 89
5.4 Pseudo Labelling As Expectation-Maximization 90
5.4.1 Problem formulation 90
5.4.2 Pseudo labels as latent variables 90
5.4.3 E-M Pseudo Labelling 91
5.4.4 On the convergence of Pseudo Labelling from the perspective of EM 93

5.5 Generalisation of Pseudo Labels via Variational Inference for Segmentation . . . 95
5.5.1 Confidence threshold as latent variable . . . 95
5.5.2 Variational E-step . . . 96
5.6 The connection between Bayesian Pseudo Label and Variational Autoencoder . . . 99
5.7 Experimental Results . . . 99
5.7.1 Data sets . . . 99
5.7.2 Baselines . . . 100
5.7.3 Training . . . 101
5.7.4 Pre processing Of the Labels For Multi-Class Segmentation 101
5.7.5 Segmentation performances . . . 101
5.7.6 Ablation studies on hyper-parameters . . . 104
5.7.7 Ablation studies on the prior of Bayesian Pseudo Labels . . 107
5.7.8 Robustness . . . 107
5.7.9 Uncertainty . . . 108
5.8 Conclusions . . . 109

6 Deep Variational Parameter Mapping: Applications of Unsupervised Representation Learning to MRI parameter estimation 110
6.1 Abstract . . . 110
6.2 Introduction . . . 111
6.3 Methods . . . 112
6.3.1 Problem formulation . . . 112
6.3.2 Probabilistic model on jointly estimating across all voxels . 113
6.3.2.1 Univariate Gaussian prior . . . 114
6.3.2.2 Gaussian Mixture Prior . . . 115
6.3.3 MRI models . . . 116
6.3.4 Implementations . . . 116
6.4 Experimental Results . . . 116
6.4.1 Baselines . . . 116

6.4.2 Data 117

6.5 Results 117

6.6 Discussion on learnt distributions of latent variables 120

6.7 Conclusion 121

7 Conclusions and Future Work 122

7.1 Conclusions 122

7.1.1 Chapter 4 122

7.1.2 Chapter 5 123

7.1.3 Chapter 6 123

7.1.4 Comparison between MisMatch and Bayesian Pseudo Labels 123

7.2 Limitations 124

7.2.1 Chapter 4 124

7.2.2 Chapter 5 124

7.2.3 Chapter 6 125

7.3 Future Work 125

7.3.1 Chapter 4 125

7.3.2 Chapter 5 125

7.3.3 Chapter 6 126

A Research Paper Declaration Form (Chapter 4) 127

B Research Paper Declaration Form (Chapter 4) 130

C Research Paper Declaration Form (Chapter 5) 133

D Research Paper Declaration Form (Chapter 5) 136

E Research Paper Declaration Form (Chapter 6) 139

Bibliography 142

List of Figures

3.1 Figure of AND, OR and XOR gate 40

3.2 One of the solutions for XOR with 2-layer perceptron, as shown in forward pass. 41

4.1 Different strategies for consistency regularisation. (a) Previous methods [1, 2, 3] use hand-crafted augmentation at input level to create predictions with different confidences. (b) Previous method [4] uses hand-crafted augmentation at feature level to create predictions with different confidences. (c) Our method end-to-end learns to create predictions with different confidences. 60

4.2 Cluster assumptions in semi supervised classification and semi supervised segmentation. (a) In classification, limited labels will cause wrong decision boundary (red straight line), where each dot is an image. (b) In classification, cluster assumption with consistency regularisation on input level perturbations at images helps to find a better decision boundary, because low density regions of images align well with the correct decision boundary. (c) In segmentation, limited labels will cause wrong decision boundary (red straight line), where each dot is a pixel. (d) In segmentation, cluster assumption with consistency regularisation on input level perturbations at pixels will not help to find a better decision boundary, because low density regions of pixels do not align with the correct decision boundary (tight boundaries between objects). 62

4.3 Strong data augmentations (e.g. shearing) change pixel-wise labels therefore they might make pixel-wise consistency regularisation not feasible for segmentation. 64

4.4 MisMatch (U-net based): decoder f_{d1} leads to dilated high confidence detection of foreground and decoder f_{d2} leads to eroded high confidence detection of foreground. The final prediction is the average between outputs of f_{d1} and f_{d2}. Any other encoder-decoder segmentation network could be used. 66

4.5 Expected calibration error [5] against accuracy in 10-fold cross-validation experiments on 50 labelled slices with CARVE. Y-axis: IoU. X-axis: ECE. Each calibration error is calculated from the gap between the confidence and accuracy for each testing image. Each data point in this figure is one testing image. The fitted 2nd order trends of our MisMatch are flatter than U-net, meaning MisMatch is more robust against the calibration error. 72

4.6 Full results of 10 fold cross-validation on CARVE. X-axis: number of labelled slices. Y-axis: IoU 73

4.7 Ablation studies on decoder architectures, cross-validation on 5 labelled slices with CARVE. MM is ours. 75

4.8 Visual results. Yellow: ground truth. Red: False Positive. Green: True Positives. Blue: False Negatives. Row 1-4: CARVE. Row 5-6: BRATS 76

4.9 Results on LA 2018 between UA-MT and MisMatch with 2 labelled cases, lr 0.01, batch 4, consistency 1 and network width 8. This further confirms that consistency regularisation on feature perturbations is more effective than consistency on input perturbations. . . . 77

4.10 Visulisation of predicted certainty of the foreground in the last positive attention shifting decoder and the last negative attention shifting decoder. We focus on the zoomed-in regions on the foreground area containing one vessel. 79

4.11 Reliability diagrams [5] from experiments on 50 labelled slices with CARVE. Blue: Confidence. Red: Accuracy. Each row is on one testing image. X-axis: bins of prediction confidences. Y-axis: accuracy. Column 1: U-net. Column 2: outputs of positive attention decoders. Column 3: outputs of negative attention decoders. Column 4: average outputs of the two decoders. The smaller the gap between the accuracy and the confidence, the better the network is calibrated. 80

5.1 Comparison between pseudo labelling approach (SegPL) and the other approaches. Pseudo labelling approaches enjoy simplicity in implementation including corresponding benefits such as scalibility and robustness. 85

5.2 Pseudo-labelling process for binary segmentation. Pseudo-label y_n' is generated using unlabelled data x_u and model with parameters from last iteration θ. Therefore, pseudo-labelling can be seen as the E-step in Expecation-Maximization. The M-step updates θ using y_n', y and data X. In our 1st implementation, namely SegPL, the threshold T is fixed for selecting the pseudo labels, which is the original pseudo labelling, as an empirical approximation of its true generalisation. In our 2nd implementation, namely SegPL-VI, the threshold T is dynamic and learnt via variational inference, which is an learnt approximation of its true generalisation. 87

5.3 Comparison between BPL and VAE in details. For BPL, only the unsupervised learning part is illustrated. 98

5.4 An example of the pre-processing of one label of BRATS. 3: enhancing tumour core. 2: tumour core containing enhancing tumour core. 1: whole tumour containing class 2 and 3. 0: healthy tissues. Different colours represent different classes. 102

5.5 Label fusion and binarized labels. Red: whole tumour including tumour core. Blue: tumour core including enhancing tumour core. Green: enhancing tumour core. Segmentation of each tumour class is a binary segmentation. 102

5.6 SegPL statistically outperforms the best performing baseline CPS when trained on 2 labelled volumes from the CARVE dataset. Each data point represents a single testing image. 103

5.7 Visual results. CARVE trained with 5 labelled volumes. Red: false positive. Green: true positive. Blue: false negative. Yellow: ground truth. GT: Ground truth. CPS: cross pseudo labels (CVPR 2021). CCT: cross consistency training (CVPR 2020). Sup: supervised training. 105

5.8 Visual results. BRATS 2018 trained with 300 labelled slices. Red: whole tumour. Green: tumour core. Blue: enhancing tumour core. GT: Ground truth. CPS: cross pseudo labels (CVPR 2021). CCT: cross consistency training (CVPR 2020). Sup: supervised training. . 106

5.9 Ablation studies on BRATS with 150 labelled slices. 106

5.10 Robustness against out-of-distribution noise. Gamma is the strength of the out-of-distribution noises. Using 2 labelled volumes from CARVE. 107

5.11 Robustness against adversarial attack. Epsilon is the strength of the FGSM[6] attack. Using 2 labelled volumes from CARVE. 107

5.12 Ablation studies on priors of 0.4, 0.5 and 0.7. Y-axis: segmentation losses. X-axis: training iterations. Red: prior = 0.5. Blue: prior = 0.4. Pink: prior = 0.7. All trained on Task 01 Brain Tumour. 107

5.13 Ablation studies on priors of 0.4, 0.5 and 0.7. Y-axis: segmentation IoUs. X-axis: training iterations. Red: prior = 0.5. Blue: prior = 0.7. Dark blue: prior = 0.5. All trained on Task 01 Brain Tumour. . 107

5.14 Learnt threshold with prior at 0.5 trained on Task 01 Brain Tumour. 109

6.1 A schematic of one of our methods with univariate Gaussian Prior. The quantitative mappings of interest are indicated by the red box. The total loss is MSE loss plus KL loss. 113

6.2 Comparisons between self-supervised voxel-wise baseline and ours (Gaussian prior) on simulated model using MSDKI. X axis: ground truth of simulated kurtosis. Y axis: prediction of kurtosis. Ours vastly outperforms the baseline in recovery of the ground truth. . . . 117

6.3 Ablation studies. We observe that both latent dimension and kl loss strength have optimal values. 118

6.4 LSQ, self-supervised, and VAE ball-stick fits to HCP dMRI subject. Our methods drastically reduces noise and better highlights anatomical features. 119

6.5 Comparisons on MS-DKI fits on HCP dMRI subject. Our approach has less obvious improvements when the MRI model is relatively simple, but doesn't hallucinate spurious anatomical features. 119

6.6 Visualisation of the posterior distribution of the latent variable in our VAE models after training on HCP data. 120

List of Tables

3.1 Truth table of XOR 40

3.2 A summary of very recent semi-supervised learning methods in medical image segmentation from MIDL, MICCAI, TMI, MedIA, CVPR and BMVC (2022-2023) 56

4.1 MisMatch (MM) vs Baselines on CARVE. Metric is Intersection over Union (IoU). 71

4.2 MisMatch (MM) vs Baselines on BRATS. Metric is Intersection over Union (IoU). 71

4.3 P-value between MM and baselines. Non-parametric Mann-Whitney U-Test. 100 labelled slices of CARVE. 74

4.4 Ablation studies on alpha value using CARVE with 5 labelled slices. 75

4.5 Ablation studies on dilation rate in 3D V-net based MisMatch using LA 2018 with 2 labelled cases, α as 1 and cutting gradients, network width 8. Metric is Dice score. 76

4.6 Ablation studies on stopping gradients in 3D V-net based MisMatch using LA 2018 with 2 labelled cases, α as 1, network width 8. Metric is Dice score. 77

4.7 Testing Results on 3D segmenting the whole tumour from Task 01 Brain Tumour from Medical Segmentation Decathlon. Training with learning rate 0.001 and 3500 epochs. Testing on $96 \times 96 \times 96$ cubes. Jac: Jaccard. HD: Hausdorff Distance. ASD: Average Surface Distance. 77

4.8 Testing Results on 3D segmenting the whole tumour from Task 01 Brain Tumour from Medical Segmentation Decathlon. Training with learning rate 0.001 and 3500 epochs. Testing on $48 \times 48 \times 96$ cubes. Jac: Jaccard. HD: Hausdorff Distance. ASD: Average Surface Distance. 78

4.9 Testing Results on 3D segmenting the whole tumour from Task 01 Brain Tumour from Medical Segmentation Decathlon. Training with learning rate 0.001 and 3500 epochs. Testing on $128 \times 128 \times 96$ cubes. Jac: Jaccard. HD: Hausdorff Distance. ASD: Average Surface Distance. 78

5.1 Hyper-parameters used across experiments. Different data might need different α. 101

5.2 Our model vs Baselines on a binary vessel segmentation task on 3D CT images of the CARVE dataset. Metric is Intersection over Union (IoU (↑) in %). Avg performance of 5 training. blue: 2nd best. red: best 104

5.3 Our model vs Baselines on multi-class tumour segmentation on 2D MRI images of BRATS 2018. Metric is Intersection over Union (IoU (↑) in %). Avg performance of 5 runs. blue: 2nd best. red: best 104

5.4 Our model vs Supervised baseline on 3D binary tumour segmentation of Task 01 Brain Tumour (BRATS 2017). Metric is Intersection over Union (IoU (↑) in %). Avg performance of models between iteration 20000 and 25000 with 1000 as the interval. red: best 104

Chapter 1

Introduction

1.1 Research problem

Deep learning has now become the de facto data-driven approach for medical image analysis in the digital era. A well-executed deep learning model requires a massive amount of data and their corresponding labels. However, acquiring labels in medicine is extremely challenging due to the high costs in both time and money. This thesis seeks the answer to the following question: *How can we train a deep learning model without sufficient labels for medical image analysis?*

1.2 Structure of the thesis

This thesis begins with Chapter 1, which outlines the structure and the contributions of the thesis. Chapter 2 presents the clinical background of this PhD, providing a high-level motivation for the necessary technical innovations. Chapter 3 reviews the machine learning foundations which are used in the subsequent technical chapters.

Chapter 4 presents a new perspective on the visual attention mechanism as learnable morphological operations at the feature level. It also introduces a new consistency regularisation on dilated feature perturbations and eroded feature perturbations for semi-supervised segmentation of medical images. Furthermore, Chapter 4 links the consistency regularisation to AI safety through a calibration analysis.

Chapter 5 revisits pseudo labelling in semi-supervised learning and provides a new formulation of pseudo labelling as the expectation maximisation algorithm. Based on this newly proposed formulation, Chapter 5 extends the original pseudo labelling towards its generalisation and presents an approximation of its generalisation using a variational inference.

Chapter 6 builds on the latest advancements in variational unsupervised clustering techniques and demonstrates its first application on model parameter estimation of MRI signals. This pioneering approach challenges the traditional assumption that all of the voxels are treated independently in MRI parameter estimations.

Chapter 7 discusses the conclusions, limitations and future work.

1.3 Publications Covered in this Thesis

The content of this thesis is based on the following publications:

- Chapter 4: **M.C. Xu**, Y.K. Zhou, C. Jin, F.J.Wilson, S.B.Blumberg, M. de Groot, D.C. Alexander, N.P. Oxtoby*, J. Jacob*, *"Learning Morphological Feature Perturbations for Calibrated Semi-Supervised Segmentation"*. **Oral** presentation at 5th International Conference on Medical Imaging with Deep Learning (**MIDL**) 2022, Switzerland.
- Chapter 4: **M.C. Xu**, Y.K. Zhou, C. Jin, M. de Groot, D.C. Alexander, N.P. Oxtoby*, J. Jacob*, *"MisMatch: Calibrated Segmentation via Consistency on Differential Morphological Feature Perturbations with Limited Labels"*. IEEE Transactions on Medical Imaging (**TMI**).
- Chapter 5: **M.C. Xu**, Y.K. Zhou, C. Jin, M. de Groot, D.C. Alexander, N.P. Oxtoby*, Y.P. Hu*, J. Jacob*, *"Bayesian Pseudo Labels: Expectation–Maximization for Robust and Efficient Semi-Supervised Segmentation"*. **Young Scientist Award Finalist** at 25th International Conference on Medical Image Computing and Image-Guided Intervention (**MICCAI**) 2022, Singapore.
- Chapter 5: **M.C. Xu**, Y.K. Zhou, C. Jin, M. de Groot, D.C. Alexander, N.P. Oxtoby*, Y.P. Hu*, J. Jacob*, *"Expectation Maximization Pseudo Labelling for Segmentation with Limited Annotations"*. Under review at MICCAI special issue of Medical Image Analysis (**MedIA**).
- Chapter 6: **M.C. Xu**, Y.K. Zhou, T. G. Allcock, K. Firoozabadi, J. Jacob, D.C.

Alexander, P. J. Slator. *"Deep Variational Parameter Mapping"*. In preparation for a technical conference submission.

1.4 Publications Not Covered in this Thesis

The following publications were conducted during the PhD but not included in the thesis. They cover a broad range of topics such as noisy labels, generative models, foundation models, and applications of AI in ophthalmology and histopathology.

- **M.C. Xu**, N.P. Oxtoby, D.C. Alexander, J. Jacob, *"Learning to Pay Attention to Mistakes"*, 31st British Machine Vision Conference (**BMVC**) 2020.
- Y.K. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyen, T. Liu, **M.C. Xu**, M. G. Lozano, A. Altmann, A. Y. Lee, E. J. Topol, A. K. Denniston, D. C. Alexander, P. A. Keane, *"A Foundation Model for Diverse and Generalisable Disease Detection from Retinal Images"*, **Nature**.
- L.Zhang*, R.Tanno*, **M.C. Xu**, J. Chen, J. Jacob, O. Ciccarelli, F. Barkhof, D.C. Alexander, *"Disentangling Human Error from the Ground Truth in Segmentation of Medical Images"*, 34th Neural Information Processing Systems (**NeurIPS**) 2020.
- A. Pakzad, **M.C. Xu**, W.K.Cheung, M.Vermant, T.Goos, L.J. de Sadeleer, S.E. Verleden, W.A.Wuyts, J.R. Hurst, J. Jacob, *"Airway measurement by refinement of synthetic images improves mortality prediction in idiopathic pulmonary fibrosis"*, 2nd International Workshop on Deep Generative Models for Medical Image Computing and Computer Assisted Interventions (**DGM4MICCAI**) 2022.
- Y.K Zhou, S. Wagner, M. Chia, A. Zhao, **M.C Xu**, R. Struyven, D.C. Alexander, P. Keane, *"AutoMorph: Automated Retinal Vascular Morphology Quantification via a Deep Learning Pipeline"*, Journal of Translational Vision Science and Technology
- **M.C. Xu***, G. Lazaridis*, S.S. Afgeh, G. Montesano, D. Garway-Heath, *"Bio-Inspired Attentive Segmentation Of Retinal OCT Imaging"*, 7th International

Workshop on Ophthalmic Medical Image Analysis at International Conference on Medical Image Computing and Computer Assisted Intervention (**OMIA**) 2020.

- Y.K. Zhou, **M.C. Xu**, Y.P. Hu, H.X. Lin, J. Jacob, P.A.Keane, D.C.Alexander, *"Learning to Address Intra-Segment Misclassification in Retinal Imaging"*, 24th International Conference on Medical Image Computing and Computer Assisted Intervention (**MICCAI**) 2021.
- C. Jin, R. Tanno, **M.C. Xu**, T. Mertzanidou, D.C.Alexander, *"Foveation for Segmentation of Mega-pixel Histology Images"*, 23th International Conference on Medical Image Computing and Computer Assisted Intervention (**MICCAI**) 2020.
- Y.K. Zhou, **M.C. Xu**, Y.P. Hu, S.B.Blumberg, A. Zhao, S.K.Wagner, P.A.Keane, D.C.Alexander, *"VAFO-Loss: VAscular Feature Optimised Loss Function for Retinal Artery/Vein Segmentation"*, Under review at Medical Image Analysis

Chapter 2

Clinical Background

In this chapter, we review the clinical background of this PhD. By understanding the clinical context, we can better appreciate the motivations for addressing the technical challenges.

2.1 Imaging Biomarkers in Drug Discovery

Pharmaceutical Research and Development is an extremely costly long-term investment with a high probability of failure. Before a new drug reaches the market, it typically goes through development at a cost of roughly 1 billion US dollars, entails more than 15 years of research work, and involves a 96% failure rate for projects [7].

A drug development journey begins at the drug discovery stage. All candidate drugs are initially validated against disease models to identify the best candidates for clinical testing. The selected potential drug candidates then progress to the experimental medicine stage, undergoing small studies to identify potential clinical endpoints. These established clinical trial endpoints will be tested under multi-centre regulatory standardisation. During the experimental medicine stage, imaging techniques can aid developers in making informed decisions amidst complex early clinical trial data. More precisely, imaging techniques can assist in several aspects, including: demonstration of target engagement; a better understanding of Pharmacokinetics—Pharmacodynamics to inform dose selection; provision of proof of pharmacology mechanism, which can also aid in assessing drug efficacy; efficient decision-making in the context of numerous trials; and examination of bioavailability and tissue exposure [8]. Alongside imaging techniques, machine learning methods can be deployed in most of the aforementioned aspects to enhance the productivity of the entire Research and Development process. In particular, the techniques being developed during this PhD could potentially assist in establishing the pharmacological mechanism during the experimental medicine stage.

2.2 Lung and Idiopathic Pulmonary Fibrosis (IPF)

Anatomical Overview The lung serves as the primary organ of the respiratory system, providing oxygen to the blood. More specifically, the respiratory system can be divided into: 1) airways, comprising the bronchi branching from the trachea and further dividing into bronchioles; and 2) lung parenchyma, consisting of bronchioles, alveoli (tiny air sacs at the end of the bronchioles), and alveolar ducts, where gas exchange takes place. Anatomically, the left lung is divided into: 1) the superior lobe; 2) the middle lobe, separated from the superior lobe by the horizontal fissure; and 3) the inferior lobe, demarcated from the middle lobe by the oblique fissure. Meanwhile, the right lung is divided into: 1) the superior lobe; and 2) the inferior lobe, which is also separated from the superior lobe by an oblique fissure.

Gas Exchange Gas exchange occurs within the alveoli and involves both the respiratory and circulatory systems. Gas exchange comprises two processes: 1) Ventilation: During inspiration, each alveolus expands its volume to intake fresh gas (high in oxygen and low in carbon dioxide) which is drawn in from the mouth via the bronchial tree. During expiration, each alveolus contracts its volume, expelling gas (low in oxygen and high in carbon dioxide) through the bronchial tree and out of the mouth; 2) Perfusion: Deoxygenated blood (low in oxygen and high in carbon dioxide), coming from the body tissues, flows into the pulmonary arteries and on to the alveolar capillaries, following the bronchi and bronchioles. After the blood passes through the alveoli and reaches the pulmonary veins, gases diffuse between the alveolar gas and blood compartments. The blood then becomes oxygen-rich and carbon dioxide-poor before it circulates back to the body tissues [9]. An impaired respiratory system can lead to many serious health problems. However, the clinical focus of this PhD is narrowed down to one specific condition, Idiopathic Pulmonary Fibrosis.

2.3 Idiopathic Pulmonary Fibrosis

History William Osler first identified the interstitial pathological patterns caused by lung diseases in the 1890s [10], with further discoveries by researchers such as Hamman and Rich who identified acute diffuse interstitial fibrosis in 1944 [10]. Alongside these anatomical discoveries, significant strides were made in physiological measures of the lungs, beginning with Borrelli's breakthrough in 1681 when he estimated the volume of air inspired in a single breath. Since then, these measures have been instrumental in diagnosing interstitial lung disease (ILD) and studying aetiology related to environment, occupation, and so on. In 1998, Katzenstein and Myers modified the pathological classification method originally proposed by Liebow and Carrington, to subtype ILDs based on histological patterns and the reparative response of the interstitium. However, the definitive universal standard for the classification and subtyping of ILDs remains an open question today. Recently, the use of high-resolution computed tomography (HRCT) has provided valuable insights into our understanding of ILDs, thanks to its ability to visualise the detailed morphological characteristics of diseased lungs.

Definition Idiopathic Pulmonary Fibrosis (IPF), historically known as Usual Interstitial Pneumonia (UIP), describes a heterogeneous series of chronic interstitial lung diseases. IPF has a higher mortality rate compared to other forms of idiopathic interstitial pneumonia (IIP). The aetiology of IPF remains unknown, hence the name 'idiopathic'. Historically, the classification of IPF was based on aetiology, a method that proved to be inaccurate because: 1) cases with different aetiologies can have similar clinical sequelae; 2) cases with similar aetiologies can have different clinical outcomes. Pathologically, IPF is characterised by thickening and stiffening of the lung tissues due to unknown causes [11]. The accumulation of such thickening eventually leads to irreversible lung scarring or fibrosis, resulting in breathing difficulties and inadequate oxygen delivery to the rest of the body. Although diagnosing and distinguishing subtypes of IPF can be challenging, it is feasible to identify disease progression in patients with fibrotic lungs based on quantification of the extent of fibrosis. Diagnosis and subtyping of IPF are challenging because: 1) UIP and non-UIP can overlap, and even experienced experts may struggle to differentiate between them; 2) the histological appearance of each IPF-affected lung varies; 3) many co-existing disorders, including desquamative interstitial pneumonia and respiratory bronchiolitis-interstitial pneumonia, are often reported in the descriptions for the same patient. Moreover, the recognition of ILD can be a suggestive sign for the future development of connective tissue diseases.

Clinical outcomes Fibroblastic foci, an avascular histologic pattern found within the interstitium, are more frequently observed in cases of Usual Interstitial Pneumonia (UIP) than in non-UIP cases [12]. In [13], it was found that the profusion of fibroblastic foci is strongly correlated with IPF. Additionally, it was concluded that patients with concurrent collagen vascular diseases have better prognoses (improved survival rates) than those without these diseases. The clinical outcomes of IPF vary between cases; however, the median survival is generally considered to be three years, while some cases might survive between five and ten years. IPF patients with differing percent-predicted baseline Forced Vital Capacity (FVC) values can have varied prognoses: 1) patients with a baseline predicted FVC less than 60% are likely to have a higher mortality rate, up to 21%; 2) patients with a baseline predicted FVC greater than 80% are likely to have a lower mortality rate, up to 6.1% [14].

Nevertheless, patients who develop acute exacerbation are unfortunately likely to progress rapidly to death. Acute exacerbation is considered a significant cause of mortality, with an in-hospital mortality rate of 56.9%, and an annual incidence between 5% to 19%. Acute exacerbation is more likely to occur among patients with increased baseline fibrosis.

Moreover, the mortality of IPF patients is also associated with systemic vascular disease. In particular, one study suggests that the risk of acute coronary syndrome, angina, and deep vein thrombosis increases before the progression of IPF.

Incidence of IPFs in UK The incidence of IPF patients in primary care in the UK grew by 35% from 2000 to 2008, with a notably higher increase in the male population in Northwest England. There are more than 5000 new cases diagnosed nationally each year, and the death rate due to IPF continues to increase [15]. Therefore, there is a pressing demand for more accurate diagnoses of IPF to enhance treatment efficacy, especially for patients at the early stages of the disease.

2.4 Treatment and Diagnosis of IPF

According to the NHS website [16], common diagnostic methods for IPF include pulmonary function tests, medical imaging methods (e.g., CT and chest X-ray), bronchoscopy, and lung biopsy. However, the prognostic signals of lung scarring in IPF are individualised and the condition can exacerbate at a rapid rate for some patients. Overall, the median survival time for IPF patients is approximately two to three years [17].

In terms of treatments, the two existing anti-fibrotic drugs, Pirfenidone/Esbriet and Nintedanib/Ofev, have been shown to be effective only in slowing down the progression of the disease, rather than halting or reversing it [18]. The outcome of this PhD might contribute to the identification of useful drug trial endpoints during Phase II/III clinical studies. This could help drug developers to more effectively develop new drugs aimed at stopping the progression of fibrosis.

Given the limitations of available medicinal treatments and the risk of acute exacerbations, the key to increasing the survival rate of IPF patients lies in early diagnosis and accurate classification of disease severity. Consequently, the measurement of IPF severity plays a central role in the severity assessment, which will be discussed in the following sections.

2.5 Clinical markers for IPF disease progression

Traditionally, the popular assessment of IPF severity is based on physiological tests (e.g., pulmonary function tests [19]) and/or radiological tests (e.g., visual classification of lung abnormalities [20]). Clinical lung function tests include spirometry, plethysmographic lung volumes, and diffusing capacity for carbon monoxide, which are performed within 24 hours of the physiological test. Using these measures, IPF severity can arguably be classified into precise stages such as ”mild”, ”moderate”, ”severe”, ”early” or ”advanced” [20]. Recently, an improved physiological measure called the 6-minute walk test has been proposed [19]. The forced vital capacity (FVC) is often used in pulmonary function tests, measuring the amount of air that can be expired from the lungs after a maximal inspiration. To assess the functionality of the alveoli, which are responsible for gas exchange, two metrics are used: 1) the transfer factor of the lung for carbon monoxide (DLco); 2) the transfer coefficient (Kco), which measures the ability of the alveoli to transport gas into the blood.

Challenges with the wide normal range of lung function tests can make it hard to distinguish the real early stages of IPF [21]. In fact, there is no standardised staging system [22] broadly accepted yet. Another limitation of lung function-based severity scoring systems has been pointed out by [23], in that lung function test results can remain stable even with disease progression. For example, as reported in [24], nearly half of the patients remained within a 10% change in forced vital capacity, while the conditions of the patients progressed.

2.6 Imaging markers for IPF disease progression

The definitive diagnosis of IPF is a combination of high-resolution computed tomography (HRCT) and surgical biopsy [22]. Since the majority of IPF patients are elderly and/or have co-morbidities, diagnosis using the non-invasive HRCT is naturally preferred by clinicians. Consequently, there has been growing interest in the community in using HRCT for studying the disease progression of IPF [20, 25, 26, 27, 28, 29].

The most typical imaging appearances of IPF are: peripheral reticular opacity at the lung bases, which may be associated with traction bronchiectasis; (almost always subpleural) honeycombing; lower lobe-predominant volume loss; and occasional irreversible ground-glass attenuation representing fine fibrosis [25]. Visual evidence has been found [27] that the extents of the abnormalities in HRCT are correlated with IPF progression. For example, reticular abnormality and some areas of ground-glass opacity can progress into honeycombing at later stages. Therefore, recent imaging-derived tools have all focused on the quantification of the aforementioned abnormalities to measure the severity of IPF. For instance, a regional texture-based quantification method was used in [28] and they found that reticular opacity is a predictor for forced vital capacity decline. Another study [30] confirmed that a quantification software called CALIPER can measure severity better than visual scoring. Later, more advanced tools such as deep learning were introduced, as seen in the study [29]. The authors used a trained binary pixel-wise classification deep learning network to segment normal and abnormal areas of IPF (e.g. reticulation, honeycombing and traction bronchiectasis). The reported score (ratio of the abnormal area against the whole lung area) has been found to correlate with the physiological evaluation results.

Through studying CT at baseline or other single time points, the community has established the fact that the pathological extent of IPF can be used as a strong indicator of poor outcome. However, literature on using CT at different time points to understand severity progression is still sparse. The potential use of serial CT to measure severity was highlighted in the study [31], where poorer outcome was visually associated with the increase of extent, which is significant as no FVC decline was identified in the same patients.

Quantification analysis derived from texture-sensitive analytic software has a primary advantage that its assessment is not affected by co-existing conditions such as emphysema or pulmonary hypertension [32] (e.g., emphysema appears in up to 50% of patients with chronic bird fancier's lung, see more details in section 2.6.1). However, issues such as inter-rater variability persist, leading to a lack of universal standards for precisely staging the disease. This issue hinders diagnosis and treatments. Other limitations and advantages are individually discussed in the following.

2.6.1 Emphysema

Emphysema coexists in about 30% of IPF patients, a condition termed as CPFE. Previous studies have led to controversial opinions on whether the co-existing damage seen in IPF, such as emphysema, has an impact on the survival rate [33]. A recent study in [34] uses HRCT to study the predictive and prognostic values of emphysema, where a calibration technique was used to eliminate the impact of disease severity at the baseline. In [34], it is found that the overall extent of total emphysema (isolated plus admixed emphysema and fibrosis) has no correlation with the ILD score. However, isolated emphysema is associated with lower disproportionately reducing gas transfer and the gas transfer coefficient; admixed emphysema preserves lung volumes (FVC) and alveolar volume. Additionally, [34] reports that admixed emphysema shows a negative correlation with traction bronchiectasis. In conclusion, [34] suggests: 1) the presence or the extent of emphysema has no additional negative impact on survival; 2) the progression of CPFE is more likely to be associated with baseline disease severity; 3) emphysema preserves lung volumes, limiting the use of FVC to study progression of CPFE. The aforementioned findings might suggest that the experimental design of some studies [35] where emphysema was included along with other low-density regions (e.g. honeycombing cysts) as the extent of the whole pathological area is not optimal.

2.6.2 Lung volume

Total lung volume is a reproducible marker for measuring lung volume loss in IPF [36]. Lung volume loss has local characteristics. The lower lobes are disproportionately reduced in volume. Although FVC has been a popular marker for measuring disease progression, changes in lung volume could be static due to the compensatory effect from emphysema, which limits the reduction of FVC decline [36]. This can lead to imprecise staging results. Additionally, the use of antifibrotic therapy could also reduce the sensitivity of using FVC decline as a measure of disease progression [36].

2.6.3 Pulmonary vascular structures

The pulmonary vessel volume (PVV) was identified as an independent predictor for mortality analysis for the first time in [37]. PVV is independent because it is not affected by the pulmonary volumetric loss from IPF progression, whereas other pathological extent-based methods might underestimate disease severity [37]. A subsequent study confirmed that PVV is better than the conventional visual score at predicting mortality [38]. Later, studies [39] [40] [41] on serial CT found that PVV increases annually by about 0.9%, suggesting that PVV can indeed be a marker to study severity progression, especially the vessels in the upper and middle zones. It has also been reported [34] that emphysema is lightly associated with a reduction in PVV.

Another promising use of vascular structures is to predict drug trial endpoints [42] [40]. In particular, the study in [42] found that by selecting patients who have a higher vessel-related structures volume than a defined threshold, the drug trial recruiters can target patients who have a more rapid FVC decline into their trial, which can in turn lead to a 26% reduction in drug trial sample size.

Due to the predictive power of pulmonary vascular structure on IPF progression and its potential use to substantially cut the cost of new drug developments, the pulmonary vascular structures are considered as the main anatomical research interest of this PhD.

2.6.4 Traction bronchiecasis

Traction bronchiectasis (fibrotic tissue pulling on the bronchi) has recently been identified as another predictive imaging marker for mortality [38]. The severity of traction bronchiectasis has been found to increase with disease progression [39]. However, the rate at which traction bronchiectasis progresses at different stages of IPF remains unknown.

2.7 Limitations of the existing quantification tool

There are three fundamental limitations to popular texture-sensitive tools such as the CALIPER software [32]: 1) limited representational power of its local histogram algorithms; 2) the domain bias in its training set; 3) inevitable domain shift issues (such as acquisition noise, machine parameters, slice thickness, and reconstruction parameters).

Firstly, CALIPER was trained on the example histogram patterns of local regions (sliding window) to recognise different pathological tissues. However, histogram-based pattern recognition techniques are not adept at distinguishing between honeycombing and emphysema, or ground glass and reticular abnormalities [32]. As for vessel recognition, CALIPER relies on the Frangi filter, and as found in preliminary experiments, a Frangi filter is likely to yield an excessive ”vesselness”. Additionally, sliding window detection has already been proven to be inferior due to its limited field of view in computer vision [43].

Secondly, CALIPER was trained on patient data from LTRC [44]. This suggests that CALIPER is severely biased towards the patient population and the training data of ”normal” lung patterns are not essentially ”healthy”. This naturally leads to a deterioration in specificity.

Thirdly, the testing data to be analysed comes from a domain which is different from the training data. This is caused by differences in acquisition and reconstruction techniques used by different centres, where data is collected on different CT machines and reconstructed with different CT algorithms. Domain shift further impairs the pattern recognition ability of a trained system on new, unseen data.

2.8 Change of the research focus of this PhD

As this work was funded by GSK, the original motivation of this PhD was to study how the increases in negative intrathoracic pressure during inspiration enlarge pulmonary arterial and venous volumes relative to expiration in idiopathic pulmonary fibrosis. GSK provided 11 cases of high contrast CT scans of patients with idiopathic pulmonary fibrosis. The original research plan of the PhD was to train a machine learning model to extract the pulmonary vascular structures for the clinical study. However, after manually reviewing the scans from GSK, we realised that the limited amount of data would hamper the significance of the clinical conclusions of the study. More importantly, we also noticed that the scans lacked paired pixel-wise labels for training machine learning models. Therefore, motivated by the practical necessity of dealing with the lack of labels in the early stage of this PhD, we decided to change the focus of this PhD from the original clinical perspective of studying idiopathic pulmonary fibrosis to a technical perspective of developing machine learning models with limited labels.

Chapter 3

Machine Learning Background Information

This chapter briefly reviews the basics of machine learning, laying the technical foundations for the following chapters. The materials used in this chapter are derived from the textbook ”Probabilistic Machine Learning: An Introduction” [45] by Kevin P. Murphy, and another textbook ”Pattern Recognition and Machine Learning” [46] by Christopher M. Bishop.

3.1 Probability theory

3.1.1 Definitions of probability and events

If one flips a coin many times, it is observed that in fair situations, the coin lands heads approximately 50% of the time. This 50% is known as the probability, which is a measure of uncertainty. Returning to the coin-flipping example, we can define an ***event*** as ”the next coin will land heads”, denoted as a binary variable A. We say $Pr(A)$ is the probability that event A is true. This $Pr(A)$ must be between 0 and 1, where 0 indicates the event definitely will not happen and 1 means that the event will definitely happen.

Probability of event A and event B both happening We can write down a joint probability of two events A and B both happening, as $Pr(A,B)$.

Probability of event A or event B happening We can also define the probability of A or B happening as $Pr(A)+Pr(B)-Pr(A,B)$. In situations where A and

B are mutually exclusive, the union distribution of A and B is $Pr(A)+Pr(B)$.

Probability of event A happening if event B happened The conditional probability of the event A happening, given another event B happened is:

$$Pr(A|B) = \frac{Pr(A,B)}{Pr(B)} \tag{3.1}$$

Independence of events In the scenario where both A and B happened, if event A and event B are conditionally independent of each other, then:

$$Pr(A,B) = Pr(A)Pr(B) \tag{3.2}$$

The above equation can be extended to include another event C if A and B are both conditional dependent on C:

$$Pr(A,B|C) = Pr(A|C)Pr(B|C) \tag{3.3}$$

3.1.2 Random variables

If X represents an unknown quantity, such as the outcome of the next coin toss, and X can take on possible values within a sample space $\mathscr{X}$, we refer to X as a random variable. For example, in the coin-flipping scenario, $\mathscr{X} = \{0,1\}$, where 0 corresponds to heads and 1 to tails. The event of the coin landing on heads is denoted as $X = 0$.

Discrete random variable If the sample space $\mathscr{X}$ contains only distinct values, we refer to X as a discrete random variable. The ***probability mass function*** represents the probability of the random variable equalling each possible value x within $\mathscr{X}$.

$$p(x) := Pr(X = x) \tag{3.4}$$

With restrictions:

$$\begin{aligned}&0 \leq p(x) \leq 1\\&\sum_{x\in\mathscr{X}} p(x) = 1\end{aligned} \tag{3.5}$$

Continuous variable If the sample space $\mathscr{X}$ contains real-valued values, we call X a continuous random variable. For continuous variable, we can partition the range of the continuous values into intervals which makes it similar to discrete random variable. When the interval goes to zero, we get the probability of X as a real value.

We can define the ***cumulative distribution function (CDF)*** of X coming from a certain range of real values, saying less than x as:

$$\begin{aligned}&P(x) := Pr(X \leq x)\\&Pr(a \leq x \leq b) = P(b) - P(a)\end{aligned} \tag{3.6}$$

The derivative of the cumulative distribution function is called ***probability density function (PDF)***. Therefore we can compute the probability of a continous variable X equal to a specific real value x, as the probability density at x multiplied with the interval dx:

$$\begin{aligned}&p(x) := \frac{dP(x)}{dx}\\&Pr(a < X \leq b) = \int_a^b p(x)dx = P(b) - P(a)\\&Pr(x \leq X \leq x + dx) \approx p(x)dx\\&Pr(X = x) = \lim_{dx\to 0} p(x)dx\end{aligned} \tag{3.7}$$

3.1.3 Derivations of Mean and Variance

Before we introduce common distributions which are used in the chapters, we need to cover two important concepts which are brought up many times in the later chapters, mean (μ) and variance (σ), of a distribution.

Mean or Expectation (μ) Expectation of the sample space of a discrete random variable is defined as:

$$\mathbb{E}[X] := \mu = \sum_{X} xp(x) \tag{3.8}$$

For continuous random variable, the Expectation is defined slightly different:

$$\mathbb{E}[X] := \mu = \int_{X} xp(x)dx \tag{3.9}$$

The property Expectation has a few intuitive properties which come very handy in applications. For example, the sum rule of the expectations of a set of random variables:

$$\mathbb{E}[\sum_{1}^{n} X_i] = \sum_{i}^{n} \mathbb{E}[X_i] \tag{3.10}$$

Similarly, the product rule of expectation applies to independent random variables as:

$$\mathbb{E}[\prod_{1}^{n} X_i] = \prod_{i}^{n} \mathbb{E}[X_i] \tag{3.11}$$

The expectation also has linearity as:

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b \tag{3.12}$$

Variance (σ) The variance measures the divergence of the distribution and it is defined based on the expectation ($\mathbb{E}[X] = \mu$) using the above Eq.3.7. Given a continuous random variable X, we can derive the variance as:

$$\begin{aligned}
\mathbb{V}[X] &:= \sigma^2 \\
&= \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{E}[(X - \mu)^2] \\
&= \underbrace{\int_{x \in \mathscr{X}} (x-\mu)^2 p(x)dx}_{definition\, of\, expectation} \\
&= \int_x x^2 p(x)dx + \mu^2 \underbrace{\int_x p(x)dx}_{=1} - 2\mu \underbrace{\int_x xp(x)dx}_{\mathbb{E}[X]} \\
&= \mathbb{E}[X^2] + \mu^2 - 2\mu\mu \\
&= \mathbb{E}[X^2] - \mu^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2
\end{aligned} \tag{3.13}$$

From Eq.3.13, the expectation of the square of $\mathscr{X}$ is the sum of the squre of mean and the square of the variance:

$$\mathbb{E}[X^2] = \mu^2 + \sigma^2 \tag{3.14}$$

People also use the standard deviation a lot which is:

$$std[X] := \sqrt{\mathbb{V}[X]} = \sigma \tag{3.15}$$

Variance has more complicated product and sum rules. For example, the sum rule of variance involves its covariance:

$$\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2Cov[X,Y] \tag{3.16}$$

The variance of the affine transformed X does not have linearity as its expectation but scaled:

$$\mathbb{V}[aX+b] = a^2\mathbb{V}[X] \tag{3.17}$$

The product rule of variance is not discussed here.

3.1.4 Distributions

The section reviews the definitions of a few distributions that are used in the later chapters.

Categorical distribution The Bernoulli distribution is used when the discrete random variable is binary in a single trial. When the discrete random variable has more than two possible values, denoted as C, the distribution is referred to as a categorical distribution. An example would be predicting the outcome of throwing a six-sided die. This distribution is used in subsequent chapters as a prior distribution in variational inference. The categorical distribution is defined as follows:

$$\begin{aligned} &Cat(x|\theta) := \sum_{c=1}^{c=C} \theta_c^{I(x=c)} \\ &\sum_{c=1}^{c=C} \theta_c = 1 \end{aligned} \tag{3.18}$$

Gaussian distribution The Gaussian distribution is defined for a continuous random variable and is widely used in this thesis. It is parameterised by two (μ and σ):

$$p(x|\mu,\sigma) := \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.19}$$

In later chapter of expectation maximization pseudo labelling, we use a special case of an univariate Gaussian called standard normal distribution, where $\mu = 0$ and $\sigma = 1$.

3.1.5 Bayesian statistics

Bayesian statistics is a common approach for modelling uncertainty using a probability distribution. Bayesian statistics play a central role in the subsequent chapter on expectation maximisation pseudo-labelling. In Bayesian statistics, we are interested in the posterior of model parameters θ, given data $\mathscr{D}$:

$$\begin{aligned} p(\theta|\mathscr{D}) &= \frac{p(\theta)p(\mathscr{D}|\theta)}{p(\mathscr{D})} \\ &= \frac{p(\theta)p(\mathscr{D}|\theta)}{\int p(\theta)p(\mathscr{D}|\theta)d\theta} \end{aligned} \tag{3.20}$$

In the above Eq.3.20, $p(\mathscr{D}|\theta)$ is known as the likelihood function, indicating that the data could have been generated from the model θ. The term $p(\theta)$ is referred to as the prior of the model parameters. The denominator $p(\mathscr{D})$ is a constant, representing the marginal likelihood of the data, or the average probability of the data. The advantage of Bayesian statistics is that, unlike frequentists who rely on running a large number of event trials, Bayesian statistics is designed to function even with just a single trial in the ideal scenario. However, it is also challenging to apply Bayesian statistics in real life because computing the marginal likelihood is a complex task, known as the intractable issue. In practice, the maximum likelihood approach is usually deployed to estimate the best model parameters, as measured by an error metric or a loss function.

3.1.6 Expectation Maximization

The Expectation Maximization (EM) algorithm is referred to extensively in Chapter 5; here, we provide a high-level introduction to the EM algorithm. The EM algorithm is an iterative method used to identify the parameters of a probabilistic model. It initially introduces a latent variable into the model, then iteratively optimises both the latent variable and the model parameters. To start with, the parameters are randomly initialised. In the E-step, we estimate the latent variable, given the existing model parameters. In the M-step, we use maximum likelihood estimation to refine the model parameters, given the latent variable from the E-step. We then repeat the E-step and M-step until convergence. More details can be found in Chapter 5.

3.2 Deep Learning

3.2.1 XOR problem

The predecessor of deep learning models is the single-layer perceptron. The perceptron model was severely criticised by Minsky and Papert as it couldn't solve non-linear decision-making problems [47]. However, the pioneers of deep learning models, often referred to as "connectionists" back in the day, discovered that by stacking two perceptrons together to make the model deeper, it was possible to solve non-linear decision-making problems such as XOR. As shown in Fig.3.1, OR and AND logic gates can be easily solved with a linear decision boundary using a perceptron, whereas XOR requires a non-linear decision boundary, which is beyond the ability of a perceptron model.

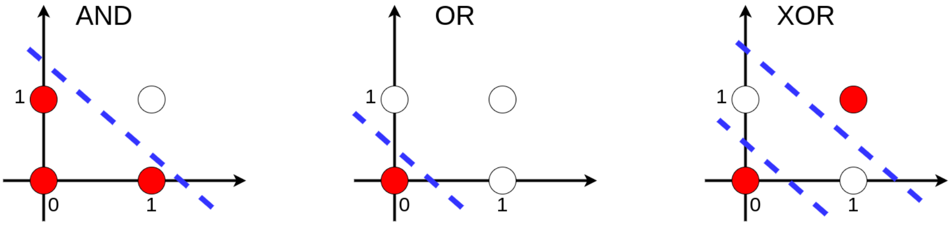

Figure 3.1: Figure of AND, OR and XOR gate

The example of how the earliest deep learning model, a two-layer perceptron, solves the XOR problem is displayed in Fig.3.1 above. The goal is to find a good set of parameters for the two-layer perceptron that aligns with the truth table in Tab.3.1.

Table 3.1: Truth table of XOR

Input 1 (x1)	Input 2 (x2)	output (y)
0	0	0
1	0	1
0	1	1
1	1	0

To recap, for the multilayer perceptron, the output of each neuron is $\phi(w^\mathsf{T} x_{\text{input}} + b)$. Here, let's assume we use a ReLU function as ϕ, which is defined as $max(x,0)$. We denote the hidden neuron as h. One possible set of weights and biases parameters could be:

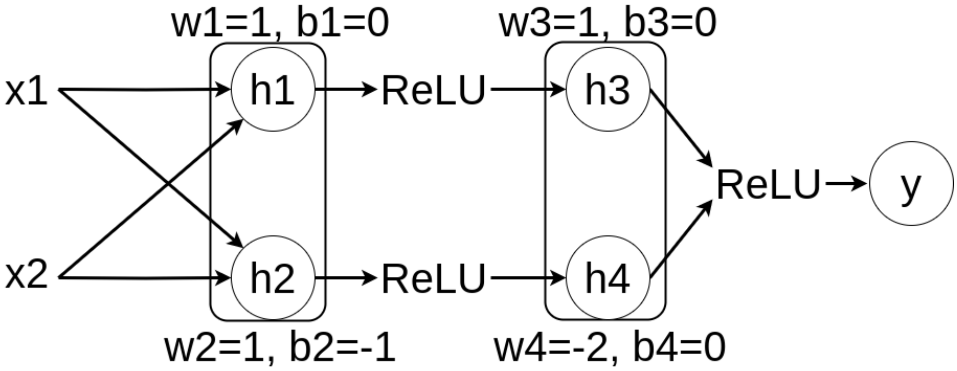

Figure 3.2: One of the solutions for XOR with 2-layer perceptron, as shown in forward pass.

For input 1 as 0 and input 2 as 0, the forward computational graph is:

$$\begin{aligned}
&x1 \longrightarrow h1 \longrightarrow h3 : max(1*0+1*0+0,0) = 0\\
&x2 \longrightarrow h2 \longrightarrow h4 : max(1*0+1*0-1,0) = 0\\
&h3,h4 \longrightarrow y : max(1*0+0+(-2)*0+0,0) = 0
\end{aligned} \tag{3.21}$$

For input 1 as 1 and input 2 as 1, the forward computational graph is:

$$\begin{aligned}
&x1 \longrightarrow h1 \longrightarrow h3 : max(1*1+1*1+0,0) = 2\\
&x2 \longrightarrow h2 \longrightarrow h4 : max(1*1+1*1-1,0) = 1\\
&h3,h4 \longrightarrow y : max(1*2+0+(-2)*1+0,0) = 0
\end{aligned} \tag{3.22}$$

For input 1 as 1 and input 2 as 0, the forward computational graph is:

$$\begin{aligned}
&x1 \longrightarrow h1 \longrightarrow h3 : max(1*1+1*0+0,0) = 1\\
&x2 \longrightarrow h2 \longrightarrow h4 : max(1*1+1*0-1,0) = 0\\
&h3,h4 \longrightarrow y : max(1*1+0+(-2)*0+0,0) = 1
\end{aligned} \tag{3.23}$$

For input 1 as 0 and input 2 as 1, the forward computational graph is:

$$\begin{aligned}
&x1 \longrightarrow h1 \longrightarrow h3 : max(1*0+1*1+0,0) = 1\\
&x2 \longrightarrow h2 \longrightarrow h4 : max(1*0+1*1-1,0) = 0\\
&h3,h4 \longrightarrow y : max(1*1+0+(-2)*0+0,0) = 1
\end{aligned} \tag{3.24}$$

Up to this point, we have demonstrated how a 2-layer perceptron can make decisions for XOR gating. Although this example might seem trivial, it illustrates that by stacking multiple layers of perceptrons together, we can create a deep model with the potential to model complex logic.

3.2.2 Brief introduction of deep learning

Deep learning is a type of machine learning model based on multilayer perceptrons, also known as neural networks. The training of a neural network requires three essential elements: 1) noise-based, gradient-based optimisation and backpropagation; 2) a cost (loss) function; 3) a model architecture composed of various layers such as normalisation layers, convolutional layers, and activation layers. An example of a three-layer deep learning model is $y = f^{(3)}(f^{(2)}(f^{(1)}(x)))$; the depth of this model is 3, and $f^{(2)}$ is the hidden layer. The term 'width' refers to the dimensionality of each layer, denoted by the total number of neurons in that layer. Each neuron models an affine mapping with one parameter called weight (w) and another called bias (b): $\phi(w^{\mathsf{T}} x_{input} + b)$, where ϕ is an activation function for non-linearity. Each layer comprises a large number of neurons. According to the universal approximation theorem, such a neural network model can approximate any arbitrary functions, with appropriate weights and biases. The approximation power of a neural network is a mapping that transforms data from the input space to the output space. There are three benefits of learning this mapping using deep learning: 1) the representation captured by a neural network is generic, which, while also achievable using kernel methods such as a Gaussian kernel or radial basis function kernel $\mathscr{K}(x, centroids) := exp(-\frac{1}{2\sigma^2}(centroids - x)^2)$, deep learning approaches generally yield superior generalisation power, especially when applied to complex problems; 2) the automatic search for task-specific features, given ambiguously precise human prior knowledge, in contrast, the computer vision community has focused on designing precise feature descriptors (e.g., SIFT) for each specific issue (e.g., edge detection) over decades. However, these methods lack transferability between tasks and require high precision of prior knowledge; 3) a significant reduction in the requirement for convexity. Traditional approaches rely on translating non-convex problems into convex problems, such as primal-dual linear programming.

3.2.3 Layers

The architecture of a neural network model depends on both the component selection and the topological design of the model. This section focuses on the components of deep learning models, also known as layers. For example, in Fig.3.2, both hidden layers belong to the same layer type, known as a fully connected layer. We focus on layers specifically for computer vision tasks and medical image analysis in 2D for illustrative simplicity, although all operations can be generalised to 3D.

Convolutional layer The convolutional layer applies a kernel with learnable weights (w) to the input (x) through a dot product operation. Here is an example of 1D convolutional layer:

$$[w \circledast x](i) = \sum_{u=0}^{L-1} w_u x_{i+u} \tag{3.25}$$

This can be generalised to 2d:

$$[w \circledast x](i,j) = \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} w_{u,v} x_{i+u,j+v} \tag{3.26}$$

Where i, j is the location of the pixel of the input and the kernel size is H, W. One detailed example of Eq.3.26 is [1]:

$$\begin{aligned} Output &= \begin{pmatrix} w_1 & w_2 \\ w_3 & w_4 \end{pmatrix} \circledast \begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \end{pmatrix} \\ &= \begin{pmatrix} (w_1x_1 + w_2x_2 + w_3x_4 + w_4x_5) & (w_1x_2 + w_2x_3 + w_3x_5 + w_4x_6) \\ (w_1x_4 + w_2x_5 + w_3x_7 + w_4x_8) & (w_1x_5 + w_2x_6 + w_3x_8 + w_4x_9) \end{pmatrix} \end{aligned} \tag{3.27}$$

For corner pixels, zero-padding is normally applied around the edges for the convolutional operations to keep the size of the output the same as the size of the input image. The kernel (W) in Eq. 3.27 applies a sliding window with a step size

[1] The original equation 14.7 on page 466 in Probabilistic Machine Learning: An Introduction contains a typo that w_4 was missing

of 1. The step size is usually referred to as the stride in the literature. However, it is also possible to skip some pixels when using a sliding window, further reducing the output size. This type of convolution is called strided convolution. There are a couple of hyperparameters of the convolutional layer, such as the size of the kernel. The size of the kernel is a trade-off between computational efficiency and performance because a larger kernel size means a larger receptive field at the cost of a larger computational burden. The term ”receptive field” refers to the size of the view of what each neuron can see in the input. In practice, researchers tend to prefer multiple small kernels rather than a single large kernel. For example, to achieve the same receptive field, two consecutive kernels of size 3 can see as much as a kernel of size 5 in a convolutional layer. Convolutional layers also have a strong relationship with linear algebra. 2D convolution is very similar to multiplication with a doubly block circulant matrix, which also explains why convolutional layers make the neural network translation equivariant.

Normalization layer: One important technique to stabilize the training of deep learning models is to use a normalization layer between convolutional layers, leading to a smoother loss function. Normalization layers calculate the mean and standard deviations from the multidimensional features and normalize the features in real-time. Depending on how the mean and standard deviations of the multidimensional features are calculated, there are different normalization layers depending on which dimension gathers statistical information.

The default normalization layer is the Batch Normalization [48] layer, which was also the first normalization layer proposed. The batch normalization layer first normalizes the feature vector (z_i) of the mini-batch with m samples, then applies an affine transformation with learnable parameters (γ and β) to the normalized features to obtain the final scaled features (z_i'):

$$\begin{aligned} \mu &= \frac{1}{m}\sum_{i=1}^{i=m} z_i \\ \sigma^2 &= \frac{1}{m}\sum_{i=1}^{i=m} (z_i - \mu)^2 \\ \hat{z}_i &= \frac{z_i - \mu}{\sqrt{\sigma^2} + \varepsilon} \\ z_i' &= \gamma \hat{z}_i + \beta \end{aligned} \tag{3.28}$$

ReLU layer Non-linear activation is typically applied after convolutional layers to provide the model with more expressiveness. The most common activation function is the rectified linear unit (ReLU):

$$ReLU(x) = max(x, 0) \tag{3.29}$$

The ReLU function acts as a high-pass filter, setting all negative inputs to zero. The biggest advantage of ReLU is its computational simplicity in both the forward and backward passes (gradients). The gradient of the ReLU function is:

$$\frac{dmax(x,0)}{dx} = \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases} \tag{3.30}$$

Later improvements have been proposed to smooth the step-wise function of the ReLU's gradient, particularly for the non-positive parts. For example, the leaky ReLU $max(\alpha x, x)$ introduces a hyperparameter α for negative values of x, ensuring that the negative parts also have gradients. Another example is the Gaussian Error Linear Unit (GELU) [49], which has become the state-of-the-art activation function in many recent Transformer [50] models, including GPTs [51].

$$\begin{aligned} GELU(x) &= x\Phi(x) \\ \Phi(x) &= P(\mathcal{N}(0,1) \leq x) \\ GELU(x) &\approx x\sigma(1.702x) \end{aligned} \tag{3.31}$$

Pooling layer Pooling layers are typically used after convolutional layers to downsample the output size. The most common pooling operation is max-pooling with a kernel size of 2 and stride of 2. The other popular class of pooling layer is the average pooling layer which uses an average operation instead of a max operation.

Fully connected layer The fully connected layer is one of the earliest layers proposed, which connects all the inputs to the neurons. In contrast to convolutional layers, which perform a local dot product through a sliding window, the fully connected layer performs a dot product operation between the entire weight matrix and the entire input matrix to generate its output.

3.2.4 Architectures

To build a neural network model, we need to assemble the layers together and create a topological architecture. Fukushima pioneered the first neural network model called neocognitron, which consisted of two types of layers: early versions of convolutional layers and pooling layers. However, the model was not trained with backpropagation. The first modern neural network, a five-layer convolutional neural network containing convolutional layers, pooling layers, and fully connected layers, was proposed by Yan LeCun in 1998 [52]. This network successfully learned to recognize written digits. In 2012, a deeper version of LeNet called AlexNet [53], proposed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, surpassed all other machine learning methods by a significant margin in the ImageNet [54] visual object recognition competition. Since 2012, deeper convolutional neural networks have been dominating the ImageNet competition with numerous advancements every year, including normalization layers and regularization techniques such as Dropout [55].

In 2017, a major breakthrough was made by Kaiming He, who introduced residual connections [56] in deep learning. The residual connections brilliantly solved the problem of vanishing gradients without incurring extra computational costs. These connections made it possible to train extremely deep neural networks, eventually leading to models that surpass human-level performance in visual object recognition. The residual connection bypasses its input to the end of the specific layer f_{layer}:

$$\hat{f_{layer}}(x) = f_{layer}(x) + x \tag{3.32}$$

In Eq.3.32, x is also called identity mapping. Following [57], we first rewrite Eq.3.32 as:

$$x_{l+1} = x_l + \mathscr{F}(x_l, W_l) \tag{3.33}$$

Where $\mathscr{F}$ is the residual function, such as two convolutional layers, and W_l is the weight of the convolutional layers in the l-th residual unit. This function is recursive. For example, the next residual connection is:

$$\begin{aligned} x_{l+2} &= x_{l+1} + \mathscr{F}(x_{l+1}, W_{l+1}) \\ &= x_l + \mathscr{F}(x_l, W_l) + \mathscr{F}(x_{l+1}, W_{l+1}) \\ &= x_{l-1} + \mathscr{F}(x_{l-1}, W_{l-1}) + \mathscr{F}(x_l, W_l) + \mathscr{F}(x_{l+1}, W_{l+1}) \\ &= ... \end{aligned} \tag{3.34}$$

The above equation says any deep layer at L can be expressed as an ensemble of the previous layers:

$$x_L = x_l + \sum_{i=l}^{i=L-1} \mathscr{F}(x_i, W_i) \tag{3.35}$$

The gradient of the loss value ($\mathbb{L}$) with respect to the i-th layer's weights parameters (W_l) is [57]:

$$\begin{aligned} &\frac{d\mathbb{L}}{dW_l} \\ &= \frac{dx_l}{dW_l}\frac{d\mathbb{L}}{dx_l} \\ &= \frac{dx_l}{dW_l}\frac{d\mathbb{L}}{dx_L}\frac{dx_L}{dx_l} \\ &= \frac{dx_l}{dW_l}\frac{d\mathbb{L}}{dx_L}\frac{d(x_l + \sum_{i=l}^{i=L-1}\mathscr{F}(x_i, W_i))}{dx_l} \\ &= \frac{dx_l}{dW_l}\frac{d\mathbb{L}}{dx_L}(1 + \sum_{i=l}^{i=L-1}\frac{d\mathscr{F}(x_i, W_i)}{dx_l}) \\ &= \frac{dx_l}{dW_l}\frac{d\mathbb{L}}{dx_L} + \frac{dx_l}{dW_l}\frac{d\mathbb{L}}{dx_L}\sum_{i=l}^{i=L-1}\frac{d\mathscr{F}(x_i, W_i)}{dx_l}) \end{aligned} \tag{3.36}$$

The gradient decomposition in Eq. 3.36 shows that the gradient at any previous layer l directly depends on the gradient at the deeper layer $\frac{d\mathcal{L}}{dx_L}$. In other words, the gradient at the deeper layer L flows back to any previous shallow layer at l through the "tunnels" of skip connections. Equation 3.36 also suggests that there is a lesser chance of vanishing gradients at the previous layer l if the network is deep with at least L layers, unless $\sum_{i=l}^{i=L-1} \frac{d\mathcal{F}(x_i,W_i)}{dx_l}$ equals -1.

3.2.5 Loss functions

The cost or loss function of deep learning typically estimates maximum likelihood from the frequentist perspective. The loss function is defined as a score to measure the degree of alignment between the model predictions and the ground truth labels.

Cross entropy One of the most common options for a cost function is cross-entropy, especially when the labels are categorical. Given the raw logits of the model output as z, the Softmax function output of z as p, and the target ground truth label y with a total of C classes, for each training data point:

$$\mathbb{L} = -ylogp \tag{3.37}$$

Where the Softmax is:

$$p_c = \frac{e_c^x}{\sum_{c=1}^{c=C} e_c^x} \tag{3.38}$$

MSE loss The mean squared error loss or l_2 loss is another very common loss function that is used in later chapters for consistency loss in semi-supervised learning.

$$\mathbb{L} = \frac{1}{N}\sum_{i=1}^{i=N}(p_i - y_i)^2 \tag{3.39}$$

3.2.6 Backpropagation

Backpropagation is the algorithm used to train a deep learning-based machine learning model. As mentioned before, a 2-layer perceptron can solve a trivial non-linear XOR problem. It was natural for the early connectionists to seek the use of more complicated perceptrons for solving more complicated non-linear pattern recognition problems. However, the problem was that there was no suitable and efficient algorithm to train such a multilayer perceptron model and find its parameters until the emergence of backpropagation. The standard method to train the parameters of the neural network is gradient-based optimization regarding the loss function, which measures how close the predictions are to the ground truth.

Mathematically, the backpropagation algorithm is the direct application of the chain rule in calculus. We show an example of the backpropagation algorithm for a two hidden-layer multilayer perceptron (MLP) model in Figure 3.2 above, from the XOR problem. If the loss function is defined as the mean squared error (MSE) loss between the prediction and the ground truth y, let's define this 2-layer MLP as a mapping of $f(x)$, where $x \in \mathbb{R}^n$ is the input and the output is $x_6 \in \mathbb{R}^m$. The final loss value as a scalar is calculated as:

$$\mathbb{L} = \frac{1}{2}||y - \phi(W_2\phi(W_1x))||_2 \tag{3.40}$$

Where W_1 are the trainable parameters (we only show weights here, biases are set up as zero for illustrative simplicity) of the 1st hidden layer, and W_2 are the trainable parameters of the 2nd hidden layer. ϕ is a ReLU function, thus parameterless. Our goal is to use backpropagation to optimize the values of the parameters W_1 and W_2.

The forward pass consists of a sequence of functions:

$$f = \underbrace{f_5}_{MSE} \circ \underbrace{f_4}_{\phi(.)} \circ \underbrace{f_3}_{W_2\times} \circ \underbrace{f_2}_{\phi(.)} \circ \underbrace{f_1}_{W_1\times} \tag{3.41}$$

The intermediate outputs of each function are:

$$\begin{aligned}
x_2 &= f_1(x) = W_1 \times x \\
x_3 &= f_2(x2) = \phi(x2) \\
x_4 &= f_3(x_3) = W_2 \times x_3 \\
x_5 &= f_4(x_4) = \phi(x_4) \\
\mathbb{L} &= f_5(x_5, y) = \frac{1}{2}||x_5 - y||_2
\end{aligned} \tag{3.42}$$

In order to update the trainable parameters W_1 and W_2 respectively, we need to calculate the gradients of the final loss scalar with respect to the target parameters:

$$\begin{aligned} \frac{d\mathbb{L}}{dW_2} &= \frac{d\mathbb{L}}{dx_5}\frac{dx_5}{dW_2} \\ \frac{d\mathbb{L}}{dW_1} &= \frac{d\mathbb{L}}{dx_5}\frac{dx_5}{dx_4}\frac{dx_4}{dx_3}\frac{dx_3}{dx_2}\frac{dx_2}{dW_1} \end{aligned} \tag{3.43}$$

We then update each layer based on the final loss scalar using the gradients. The details of how to calculate the gradients efficiently are not included here. Backpropagation grants the model the ability to learn, differing from the way biological intelligent systems operate. Combined with later hardware revolutions, featuring much more powerful computing chips, the AI revolution finally kicked off in 2012 after the deep learning model AlexNet won the ImageNet competition for object recognition.

3.2.7 Stochastic Gradient Descent

In deep learning, the most commonly used optimization technique is called stochastic gradient descent (SGD), which usually operates in batches. Although SGD does not guarantee global minima of solutions, it introduces stochastic noise, which is a key factor leading to the convergence of the algorithm. It is worth mentioning that SGD relies on suitable initialization of network parameters. The goal here is to minimize the scalar loss value with respect to the model parameters:

$$\mathbb{L} = \mathbb{E}_{x \sim q(x)}[loss\,function(model, x)] \tag{3.44}$$

Where $q(x)$ is the distribution of the training images. However, computing over the entire $q(x)$ is too challenging. In practice, we aim to estimate the gradient of $\frac{d\mathbb{L}}{dx}$ by randomly sampling a small subset from $q(x)$ at every iteration t:

$$\mathbb{L}_t = loss\,function(model, x_t) \tag{3.45}$$

We then calculate the gradient of $\mathbb{L}_t$ instead and update the model parameters θ at each iteration t as:

$$\theta_{t+1} = \theta_t - gamma \nabla \mathbb{L}_t(\theta_t, x_t) \tag{3.46}$$

Where γ is the step size of optimization, and it is normally set up as a hyperparameter called learning rate. The learning rate γ is the most crucial hyperparameter in network training, and it has been widely observed that different learning rate annealing schedules could aid convergence.

Adam [58] A variant of the SGD algorithm is now the default training algorithm for deep learning models, called ADAM (adaptive moment estimation). ADAM is a combination of adaptive gradient methods and momentum methods. ADAM first computes the 1st and 2nd order moments of gradients using exponentially weighted moving averages before updating the model parameters:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1)\nabla \mathbb{L}_t(\theta_t, x_t) \\ s_t &= \beta_2 s_{t-1} + (1-\beta_2)\nabla^2 \mathbb{L}_t(\theta_t, x_t) \\ \theta_{t+1} &= \theta_t - gamma \frac{1}{\sqrt{s_t} + \varepsilon} m_t \end{aligned} \tag{3.47}$$

Bias correction is normally used for the momentum in the beginning of the training:

$$\begin{aligned} \hat{m}_t &= m_t/(1-\beta_1^t) \\ \hat{s}_t &= s_t/(1-\beta_2^t) \end{aligned} \tag{3.48}$$

3.2.8 Training tricks

At its core, deep learning is an engineering science; therefore, its performance depends on a lot of heuristic tricks. For example, the initialization of the weights and other parameters also plays a crucial role in training. The initial weights are typically sampled from a Gaussian distribution, and the parameters of the Gaussian distribution might affect the initialization effect. Regularisation is also vital to prevent overfitting. Commonly used techniques include dropout layers, which randomly disable connections between neurons with a hyperparameter probability.

Weight decay is another typical technique to prevent overfitting, which is equivalent to l_2 regularization in ridge regression.

3.3 Deep Learning for Medical Image Segmentation

Segmentation: Given an unseen pixel (x) from a medical image, the goal is to determine the label (y) for that x. This is typically achieved using a maximum likelihood estimation approach, implemented as a supervised learning algorithm trained on seen images with paired pixel-wise labels. Essentially, recognising the labels of the pixels involves estimating a probability distribution $p(y|x;\theta)$, where θ represents the model parameters. The default task is binary pixel-wise segmentation, where the probability of y is 1 when the algorithm determines x to be in the foreground class, and the probability of y is 0 for the background class. Thus, the sum of the two candidate probabilities is 1. To assign labels more effectively, the probability distribution is squashed into the interval between 0 and 1 using logistic regression (Sigmoid function): $p(y=1|x;\theta)=\sigma(\theta^{\top}x)$. It should be noted that the normalization function is typically Softmax in a multi-class scenario.

Challenges: In recent decades, the volume of medical imaging data has become ”big”. As a result, state-of-the-art data-driven methods have been introduced to the medical imaging domain to address important issues such as more efficient and accurate diagnosis. However, the amount of available medical imaging data is still significantly smaller than that of natural imaging data. This is because: 1) data acquisition in the medical domain is expensive and requires medical devices; 2) acquiring ground truth data is very expensive as it requires expertise from medical professionals. These factors pose several technical challenges that need to be addressed by the medical imaging community, including: 1) the issue of label scarcity; 2) the issue of noisy labels and inter/intra-rater variability; 3) efficient learning of volumetric data.

Nowadays, computer-aided diagnosis relies on pixel/voxel-wise segmentation of medical images, where each pixel or voxel is classified into a category belonging to different classes such as background healthy tissue, foreground pathological tissue, or different types of soft tissues or bones. The decision-making process is based on the learned features of the models. Since the renaissance of deep learning due to AlexNet [59], numerous deep learning-based segmentation models have been proposed. Among them, U-Net [60] has emerged as the most widely adopted model due to its superior performance.

3.3.1 Supervised Learning

Early deep learning models in medical image segmentation were based on convolutional networks used for image classification methods. Due to the use of fully connected layers for classification in those networks, the sliding window method had to be enabled to classify each patch of a volume, as demonstrated by the classical method DeepMedic [61]. In DeepMedic, the authors developed a two-stream approach, with each stream taking inputs at different resolutions. The fully convolutional network was the first model to replace the fully connected layer with a convolutional layer for end-to-end pixel-wise classification. U-Net [62] inherited the merits of the originally proposed fully convolutional network (FCN) [43]. Like FCN, U-Net can perform classification at each pixel regardless of the resolution of input images. The success of U-Net comes from the intensive use of skip-connections between its encoder and decoder. These skip-connections have the following advantages: 1) the model makes decisions based on fused low-level and high-level features, with low-level features being especially important in greyscale medical imaging. Previous approaches might not have effectively utilized low-level features; 2) reusing low-level features is an efficient learning strategy, particularly given the limited availability of labelled training data. Additionally, this feature reuse can be seen as a form of regularization that reduces overfitting; 3) skip-connections, in general, aid optimization, especially in reducing the issue of exploding gradients and smoothing the loss surface. Another important factor of U-Net is its symmetric design, where the decoder has a capacity as large as the encoder. Previous methods, on the other hand, employed very lightweight decoders. Sequence-modelling-inspired architectures such as Transformers have also gained tremendous research interest in the community. However, as of early 2023, there is still no concrete evidence that Transformers have fully surpassed U-Net in medical image segmentation tasks, judging from recent results from MICCAI segmentation competitions. Another important aspect to consider is that U-Net has significantly lower computational requirements compared to Transformers, which might be another reason why U-Net remains the most popular segmentation framework to date. Surprisingly, U-Net has also shown its potential in other recent advancements, such as generative diffusion models and GANs. The potential improvements and applications of U-Net architecture still remain one of the most active research fields in neural network architectures.

State-of-the-Art: In the segmentation of medical images, a recent framework called nnU-Net [63], standing for ”not-new U-Net,” consistently outperformed many complicated models across different tasks. The most striking uniqueness of nnU-Net is that it dedicates most of its efforts to configuring a domain-specific data preprocessing pipeline and implementing good machine learning practices. In terms of modelling, nnU-Net uses the standard U-Net [62]. The standardized data preprocessing pipeline drastically improves the quality of training data, resulting in its leading performance on several leaderboards of medical image segmentation competitions. For example, in all of the MICCAI challenges in 2021, 5 out of 7 were based on nnU-Net. The inspiring success of nnU-Net breaks the medical imaging community's stereotype that more advanced models outweigh more careful data engineering work in achieving good performance. nnU-Net verifies that good input is as important as a good model.

3.3.2 Semi-Supervised Learning

Semi-supervised learning (SSL) is a branch of representation learning in which the training data contains both labelled and unlabelled data. SSL training, therefore, consists of two parts: supervised learning on labelled data and self-training on unlabelled data. Most popular semi-supervised learning methods have the same learning objective, which is based on maximising the mutual information between the input and the output.

3.3.2.1 SSL in Classification

Popular classes of common SSL methods have been compared in a benchmark study [64]. A direct application of the smoothness assumption is called label propagation, which propagates labels to unlabelled data based on the similarity between labelled and unlabelled data [65]. However, constructing similarity graphs for label propagation involves computationally heavy Laplacian matrices, leading to scalability issues. Another common method is entropy minimisation, which aims to drive models to achieve low-entropy predictions on unlabelled data [66] [67]. One drawback of entropy minimisation is the risk of overfitting, which can lead to incorrect decision boundaries for data points in low-density regions (see Appendix E in [64]). Other attempts include generative models, such as the one proposed in [68], which combines GANs in training but suffers from unstable training. The state-of-the-art methods are dominated by consistency regularisation methods, as they are easy to use and effective across different tasks. Among the consistency regularisation methods, Mean-Teacher [1] is the most representative example. It involves two identical models that are fed inputs augmented with different Gaussian noises. The first model learns to match the target output of the second model, while the second model uses an exponentially moving average of the parameters from the first model. One of the state-of-the-art SSL methods [69] [2] combines entropy minimisation and consistency regularisation.

3.3.2.2 SSL in Segmentation

In semi-supervised image segmentation, consistency regularisation is commonly used [70] [71] [72] [73] [74] [3], where different data augmentation techniques are applied at the input level. Another related work [75] forces the model to learn rotation-invariant predictions. In addition to augmentation at the input level, feature-level augmentation has gained popularity in consistency-based SSL segmentation [4, 76]. There have also been attempts to create perturbations using dual network branches [77] [78]. However, in contrast to [77] and [78], the perturbations we use are also learned via the network itself. Apart from consistency regularisation methods in medical imaging, other attempts have been made, including the use of generative models to create pseudo data points for training [79] [80], as well as the use of different auxiliary tasks as regularisation [81] [82].

A summary of recently proposed semi-supervised medical image segmentation methods can be found in Table 3.2.

References	Dimension	Modality	Datasets	Highlights
Consistency Regularisation				
MIDL2023 [83]	3D	CT/MRI	ACDC/AMOS/BraTS	Multi-scale
MedIA2023 [84]	3D	MRI	ACDC/Prostate/MMWHS	Contrastive loss for consistency
TMI2023 [85]	3D	MRI	ACDC/Prostate/PROMISE	Consistency on adversarial noise
MIDL2022 [86]	2D	CT/MRI	CARVE/BraTS	Consistency on morphological feature perturbations
MIDL2022 [87]	3D	CT	ACDC	Consistency on two different networks
TMI2022 [88]	3D	CT	COVID-SemiSeg/SEG-C19/LIDC	SwapMix data augmentation
BMVC2022 [89]	2D	MRI	ACDC	MixUp data augmentation
TMI2022 [90]	2D/3D	MRI/CT	LA/MS-CMR/Hippocampus(Vandervilt).	Consistencies on contextual and structural features
TMI2022 [91]	2D	Fundus	DRIVE	Combined with domain adaptation
MICCAI2022 [92]	2D	Histological	MoNuSeg	Consistency for each scale
MICCAI2022 [92]	2D	OCT	SEG/UKBB	Consistency on geometrical boundaries from different methods
MICCAI2022 [93]	2D	Endoscopic	Kvasir/CVC-ClinicDB/EndoScene/ETIS-Larib-Polyp-CVC-ColonDB	Temporal consistency on adjacent frames
TMI2022 [94]	2D	Fundus	SEG/UKBB	Geometrical constraints on two branches of Graph Neural Network
MedIA2022 [95]	3D	MRI	BraTS/Prancreas-NIH/NPC-MRI	Consistency across different scales
TMI2022 [96]	2D	Histological	AC3/AC4/CREMI/Kasthuri15	Mean-Teacher
MICCAI2022 [97]	2D	MRI		Consistency on different models outputs
MedIA2022 [98]	3D	MRI	PROMISE12/ACDC	Teacher model uses dropouts, student model uses signed distance function
MedIA2022 [99]	3D	Ultrasound	EchoNet-Dynamic/CAMUS	bi-directional spatiotemporal features fusion models
Pseudo Labelling				
TPAMI2023 [100]	3D	CT	HUST-COVID	Multi-task
MICCAI2022 [101]	3D	CT/MRI	CARVE/BraTS	A new Bayesian formulation of pseudo labels
TPAMI2023 [102]	3D	CT/MRI	ACDC/MMWHS/REFUGE	A new formulation of pseudo labels via risk estimation
MICCAI2022 [103]	3D	MRI	LA/Pancreas	Uncertainty-aware
MICCAI2022 [104]	3D	CT	Med. Seg. Decathlon	Active learning framework with Human feedback
MICCAI2022 [105]	3D	MRI	ACDC	Extra local contextural constrinats
MICCAI2022 [106]	2D	OCT	CORN-1/CCM30/BJH	Active learning framework
MICCAI2022 [107]	2D	MRI	ACDC/MM-WHS	Fuzzy fusion for pseudo labelling
CVPR2022 [108]	2D	Histological	DSB/MoNuSeg	Contrastive learning on features from different patches
MICCAI2022 [109]	2D	MRI	BraTS	Combined with domain adaptation
Generative Modelling				
TMI2022 [110]	2D	CT/MRI	LiST/ISIC/LA	GAN as regularisation
TMI2022 [111]	3D	CT	COVID-19-20/Mosmed-1110	Generate pathological areas highlighed by pseudo labels
TMI2022 [112]	3D	Fundus/OCT	Private	Adversarial training to tell the prediction apart from label
TMI2022 [113]	3D	NIR	VESSEL-NIR	LSTM and GAN for generating sequential images
CVPR2022 [114]	3D	CT	KiTS19/AtrialSegChallenge	Generate the input volumes as regularisation
TMI2022 [115]	3D	MRI	MIDAS	Segmentation base model is a Transformer
Hybrid				
CVPR2023 [116]	2.5D	CT	LA/KiTS19/LiTS	Two branches, uncertainty-aware
CVPR2023 [117]	3D	CT/Hist.	ACDC/KiTS19/CRAG	teacher-student model
CVPR2023 [118]	3D	CT/MRI	LA/NIH Pancreas	teacher-student model, pseudo labels fusion
MedIA2022 [119]	2D	CT/MRI	Hecktor/BraTS	Novel area similarity contrastive loss

Table 3.2: A summary of very recent semi-supervised learning methods in medical image segmentation from MIDL, MICCAI, TMI, MedIA, CVPR and BMVC (2022-2023)

3.3.2.3 Mutual Information and entropy minimisation

Many semi-supervised learning methods aim to maximize the mutual information between the input and output. We can view the learning of a model as a mapping between the high-dimensional imaging input and the low-dimensional categorical output. It is natural to assume that the information is preserved between the input and output. This assumption is particularly important in scenarios where labels are not available, and thus we can formulate an objective function for unlabelled data. As described by Birdle [120], such an objective function can intuitively aim to maximize the mutual information between the input and output of the unlabelled data. The mutual information $I(y;x)$ between the input x and the output y can be defined as [120] based on the definition of forward KL divergence ($kl(p(x)||q(x)) = \int p(x)\frac{p(x)}{q(x)}dx$):

$$
\begin{aligned}
I(y;x) &= KL(p(x,y)||p(x)p(y)) \\
&= \int\int p(y,x)log\frac{p(y,x)}{p(y)p(x)}dydx \\
&= \int\int p(y|x)p(x)log\frac{p(y,x)}{p(y)p(x)}dydx \\
&= \underbrace{\int p(x)dx}_{\int dxp(x)(.):\mathbb{E}_x} \int p(y|x)log\frac{p(y|x)}{p(y)}dy
\end{aligned}
\tag{3.49}
$$

Now let's say if the y contains C total classes for a classification task, then [120]:

$$
\begin{aligned}
I(y;x) &= \mathbb{E}_x[\sum_{i=1}^{i=C} p(y_i|x)log\frac{p(y_i|x)}{\mathbb{E}_x(p(y_i)|x)}] \\
&= \mathbb{E}_x[\sum_{i=1}^{i=C} p(y_i|x)log(y_i|x)] - \sum_{i=1}^{i=C} \mathbb{E}_x[p(y_i|x)log\mathbb{E}_x[p(y_i|x)]] \\
&= -\mathbb{E}_x[\mathbb{H}(y)] + \mathbb{H}(\mathbb{E}_x[y]) \\
&= \mathbb{H}(\bar{y}) - \mathbb{H}\bar{(}y)
\end{aligned}
\tag{3.50}
$$

The above loss function [120] is the entropy of the average of the outputs minus the average of the entropy of the outputs. To maximise the above loss function, we need to maximise the 1st term $\mathbb{H}(\bar{y})$. For each training data, it should be evenly distributed across all classes, which corresponds to entropy minimisation. Meanwhile, the second term should be minimised. For each training data, it should have a maximum probability for one class compared to all other classes, which can be implemented using pseudo-labelling.

Chapter 4

MisMatch: Calibrated Segmentation via Consistency on Differential Morphological Feature Perturbations with Limited Labels

This chapter describes differential morphological operations via neural networks and a novel consistency regularisation framework for semi-supervised segmentation. The current form of this chapter has been published in IEEE Transactions on Medical Imaging. A shorter version of this chapter has been presented as an oral presentation (top 11.6 %) at the 5th International Conference on Medical Imaging with Deep Learning (MIDL) 2022. I conceived the idea, implemented the code, performed the experiments and wrote the draft for the manuscript; my colleague Yukun provided feedback on experiments design; all of the co-authors contributed to the writing of the manuscript.

4.1 Abstract

Semi-supervised learning (SSL) is a promising machine learning paradigm to address the ubiquitous issue of label scarcity in medical imaging. The state-of-the-art SSL methods in image classification utilise consistency regularisation to learn unlabelled predictions which are invariant to input level perturbations. However,

image level perturbations violate the cluster assumption in the setting of segmentation. Moreover, existing image level perturbations are hand-crafted which could be sub-optimal. In this paper, we propose MisMatch, a semi-supervised segmentation framework based on the consistency between paired predictions which are derived from two differently learnt morphological feature perturbations. MisMatch consists of an encoder and two decoders. One decoder learns positive attention for foreground on unlabelled data thereby generating dilated features of foreground. The other decoder learns negative attention for foreground on the same unlabelled data thereby generating eroded features of foreground. We normalise the paired predictions of the decoders, along the batch dimension. A consistency regularisation is then applied between the normalised paired predictions of the decoders. We evaluate MisMatch on four different tasks. Firstly, we develop a 2D U-net based MisMatch framework and perform extensive cross-validation on a CT-based pulmonary vessel segmentation task and show that MisMatch statistically outperforms state-of-the-art semi-supervised methods. Secondly, we show that 2D MisMatch outperforms state-of-the-art methods on an MRI-based brain tumour segmentation task. We then further confirm that 3D V-net based MisMatch outperforms its 3D counterpart based on consistency regularisation with input level perturbations, on two different tasks including, left atrium segmentation from 3D CT images and whole brain tumour segmentation from 3D MRI images. Lastly, we find that the performance improvement of MisMatch over the baseline might originate from its better calibration. This also implies that our proposed AI system makes safer decisions than the previous methods.

4.2 Introduction

Training of deep learning models requires a large amount of labelled data. However, in applications such as in medical image analysis, anatomic/pathologic labels are prohibitively expensive and time-consuming to obtain, with the result that label scarcity is almost inevitable. Advances in the medical image analysis field requires the development of label efficient deep learning methods and accordingly, semi-supervised learning (SSL) has become a major research interest within the community. Among the myriad of SSL methods used, consistency regularisation based methods have achieved the state-of-the art in classification [1, 2, 69, 121], thus we focus on this genre in this paper.

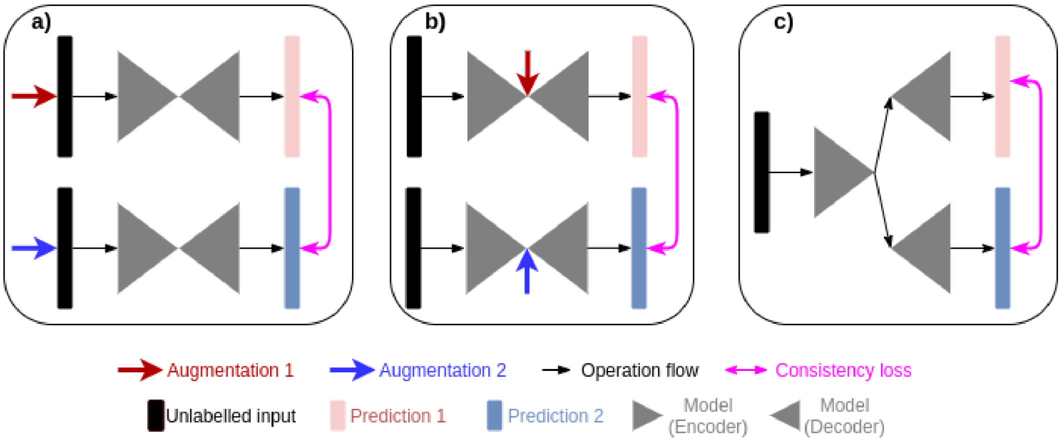

Figure 4.1: Different strategies for consistency regularisation. (a) Previous methods [1, 2, 3] use hand-crafted augmentation at input level to create predictions with different confidences. (b) Previous method [4] uses hand-crafted augmentation at feature level to create predictions with different confidences. (c) Our method end-to-end learns to create predictions with different confidences.

Existing consistency regularisation methods [1, 2, 69, 121, 4, 76, 3, 75] are mainly focusing on producing predictions which are invariant against different input level perturbations. In other words, we can interpret that consistency regularisation methods aim at training networks which generate augmentation invariant predictions. For example, if we apply weak augmentation such as flipping on an input image, the model will assign a high probability of this image belonging to its correct label, hence, the prediction of the weakly augmented image is with high confidence; if we apply strong augmentation such as rotation on an input image, then the testing is much more difficult and the model might assign a low probability of this image to its correct label, therefore, such a prediction of a strongly augmented image is with low confidence. A consistency regularisation is enforced to align the paired predictions. The relationship between consistency regularisation and augmentation invariant predictions imply that such networks should be having better calibration, which will be empircally verified in section 4.8. However, data augmentation techniques used in existing semi-supervised learning are typically hand-crafted which might be sub-optimal. Practically, such augmentation techniques are not adaptive across pixels which may be problematic as spatial correlations amongst pixels are crucial for segmentation, e.g. neighbouring pixels might belong to the same category. Most importantly, direct adaption of input level perturbations in segmentation violates the cluster assumption which is the foundation of semi supervised learning, we will explain this issue further in later section 4.3.

In this paper, we propose an end-to-end learning framework to generate predictions with different confidences. In order to change prediction confidences at a pixel-wise level in a realistic way, we use two different attention mechanisms to respectively dilate and erode foreground features which correspond to the areas of "ground truth". A preliminary version of this manuscript has been presented at MIDL 2022 [86]. Comparing to the previous MIDL version, we now included extra experiments on two 3D data sets using a different base network; a more detailed explanation of the motivation; a more principled method section under the guidance of the theory of effective receptive field. The code is here: https://github.com/moucheng2017/MisMatchSSL. Our contributions are summarised as:

- We provide an intuition of the relationship between consistency regularisation and semi-supervised learning, and why consistency regularisation with data augmentation wouldn't work well in segmentation.
- We propose a framework called MisMatch for semi supervised segmentation, by combining differential morphological feature perturbations with consistency regularisation.
- We discovered that our consistency regularisation improves model calibration, leading to safer AI deployment for medicine.
- We intensively evaluated our framework on four medical applications including: 1) 2D segmentation of lung vessel of CT images; 2) 2D segmentation

of brain tumour of MR images; 3) 3D segmentation of left atrium of MR images; 4) 3D segmentation of whole tumour from MRI images. We conclude that our consistency regularisation on feature perturbations is more effective than consistency on input level perturbations.

4.3 Motivations

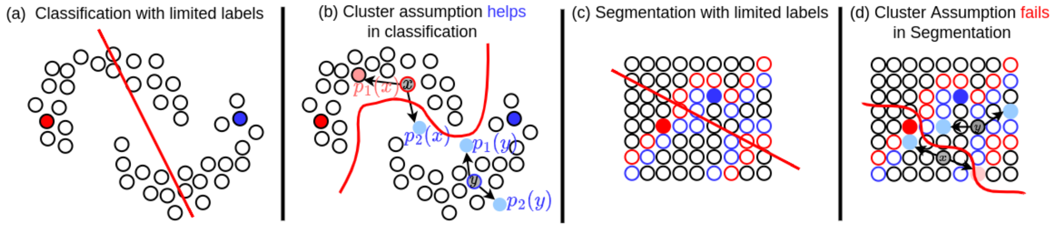
(a) Classification with limited labels
(b) Cluster assumption helps in classification
$p_1(x)$
x
$p_2(x)$
$p_1(y)$
y
$p_2(y)$
(c) Segmentation with limited labels
(d) Cluster Assumption fails in Segmentation
x
y

Figure 4.2: Cluster assumptions in semi supervised classification and semi supervised segmentation. (a) In classification, limited labels will cause wrong decision boundary (red straight line), where each dot is an image. (b) In classification, cluster assumption with consistency regularisation on input level perturbations at images helps to find a better decision boundary, because low density regions of images align well with the correct decision boundary. (c) In segmentation, limited labels will cause wrong decision boundary (red straight line), where each dot is a pixel. (d) In segmentation, cluster assumption with consistency regularisation on input level perturbations at pixels will not help to find a better decision boundary, because low density regions of pixels do not align with the correct decision boundary (tight boundaries between objects).

Cluster assumption In this section, we will explain the cluster assumption for semi-supervised classification and how it is violated if we straightforwardly transfer existing consistency regularisation methods from classification to segmentation. The cluster assumption is a variant of the smoothness assumption. The smoothness assumption states that if two data points (x_1 and x_2) are adjacent to each other, their outputs or labels (y_1 and y_2) should also be close to each other. The cluster assumption directly derives from the smoothness assumption, for example, if there is a dense population of data points in a space, then it is highly likely that cthe luster of those densely neighbouring data points are in the same class. In other words, the cluster assumptions implies there exists low density regions among different classes or different clusters of data points and the correct decision boundary should lie at the low-density regions. Equivalently, the key is to find the low-density regions which leads to the rightful decision boundary.

Consistency with Data Augmentation in Classification We start with a classical two moon example to explain how consistency regularisation with data augmentation works in semi-supervised classification. Each moon represents a class and each dot represents an image for semi-supervised classification. As shown in the two moons example in Fig. 4.2(a), if there are very limited labelled data points such as two data points, any decision boundary between the two labelled data points is possible, for example, the examplar decision boundary shown in Fig. 4.2(a) can wrongly classify half of the data points. The two moon example in Fig. 4.2(a) and (b) is also a perfect example for cluster assumption that the low density region between the two moons can separate the two moons from each other. In Fig. 4.2(b), let's focus on the two images x and y which are from the upper moon class and lower moon class respectively. If we apply two random augmentations (directional arrows in Fig. 4.2(b)) on the images, we will get $p_1(x)$ and $p_2(x)$ from x, $p_1(y)$ and $p_2(y)$ from y. Since x is closer to the low-density region, the augmented x could cross the decision boundary thereby $p_2(x)$ could be wrongly classified as the lower moon class, meanwhile, $p_1(x)$ still stays in the cluster of upper moon class. In this case, $p_1(x)! = p_2(x)$ although they are derived from the same data point x. The difference between $p_1(x)$ and $p_2(x)$ will be more than 0 which can be back-propagated to optimise the model parameters. On the contrary, the image y is closer to the centre of the cluster of lower moon class, that $p_1(y)$ and $p_2(y)$ are the same, resulting in 0 differences which does not affect the model parameters. Hence, it is easy to tell that the consistency regularisation with data augmentation makes the model parameters sensitive to the images closer to the low-density regions. This property will naturally drive the model to locate the low-density regions which happen to be the correct decision boundary.

Consistency with Data Augmentation in Segmentation However, consistency regularisation with data augmentation will have clear limitations in segmentation. In segmentation, as shown in Fig. 4.2(c), now we have each dot as a pixel and all of the pixels are densely distributed across the image space. In Fig. 4.2(c) and (d), we highlight the object boundary with continuous red and blue dots along the two sides of the boundary respectively. As there are hardly low-density regions between objects, it becomes hard to align the objects boundaries with low-density regions. If we have only two labelled pixels from each class, we will not be able to locate the correct decision boundary as illustrated in Fig. 4.2(c). If we apply two different augmentations on x and y with consistency regularisation as shown in Fig. 4.2(d), although the model can still locate the pixels which are sensitive to the consistency regularisation, due to the lack of clear low-density regions, the model will not correctly locate the right decision boundaries.

Practical Limitations of Strong Data Augmentations in Segmentation Common strong data augmentation techniques typically distort the spatial characterisation of the objects such as shearing. As shown in Fig.4.3, the image-wise label

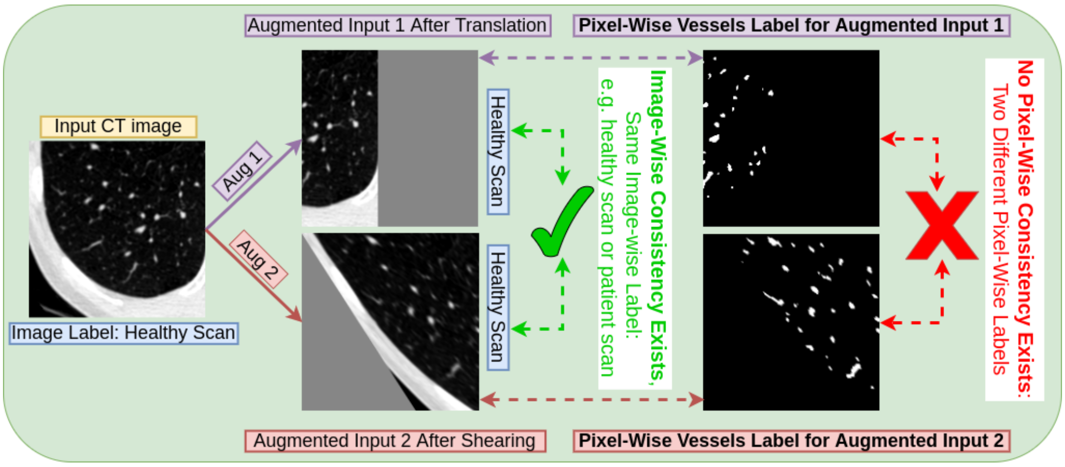
Augmented Input 1 After Translation
Pixel-Wise Vessels Label for Augmented Input 1
Input CT image
Aug 1
Aug 2
Healthy Scan
Healthy Scan
Image-Wise Consistency Exists,
Same Image-wise Label:
e.g. healthy scan or patient scan
No Pixel-Wise Consistency Exists:
Two Different Pixel-Wise Labels
Image Label: Healthy Scan
Augmented Input 2 After Shearing
Pixel-Wise Vessels Label for Augmented Input 2

Figure 4.3: Strong data augmentations (e.g. shearing) change pixel-wise labels therefore they might make pixel-wise consistency regularisation not feasible for segmentation.

stay the same, regardless of the data augmentation is applied. However, strong data augmentation will modify the pixel-wise labels, leading to difficulty of applying consistency regularisation at pixel-wise if two different strong data augmentations are applied on the same image. To avoid this practical issue, specific strong data augmentation such as CutMix was chosen in order to use consistency regularisation in segmentation [3]. In our paper, we propose an alternative solution. We use augmentation at the feature level in lieu of augmentation of the data level, to completely avoid this practical issue.

Proposal Although the low-density regions do not align with the objects boundaries anymore, the evidence in [4, 3] suggests that the low-density regions actually align well with the objects boundaries in the feature space. This means that it might be possible to use consistency regularisation on the predictions which are invariant to feature perturbations to identify the correct decision boundaries in segmentation. This directly inspired us to focus on feature perturbations in our work that we want to design learnable feature perturbations which are realistic and semantically meaningful. More specifically, we decide to apply morphological-alike perturbations on the features. In the following sections, we show how to use inductive biases of neural network topology to ask the networks to end-to-end learn morphological feature perturbations.

4.4 Methods

4.4.1 Background: ERF and the foreground

Effective Receptive Field We introduce how to control the size of the foreground features by controlling the effective receptive field (ERF). The ERF [122] measures the size of the effective area at the centre of receptive field and it impacts the most on the prediction confidence of the central pixel of the receptive field, which should overlap with the foreground objects with the highest confidence at the foreground central pixel. If we want to apply morphological operations on features of foreground objects, equivalently, we need to adjust the ERF on the foreground. As found in [122], a larger ERF means the model can effectively take a larger area of the image into account during inference of decision making, resulting in higher prediction confidence at the centre, meanwhile, a smaller ERF leads to less confident prediction on the central pixel due to the lack of visual information of neighbouring pixels. More importantly, ERF is highly affected by the network architecture. In particular, the dilated convolutional layer can increase the ERF to an extent dependent on the dilation rate [122]. Skip-connections conversely can shrink the ERF, though the extent of this effect is as yet unknown [122]. We are therefore inspired by [122] to design a network to control the ERF, in order to deliberately change the prediction confidence to morph the foreground features.

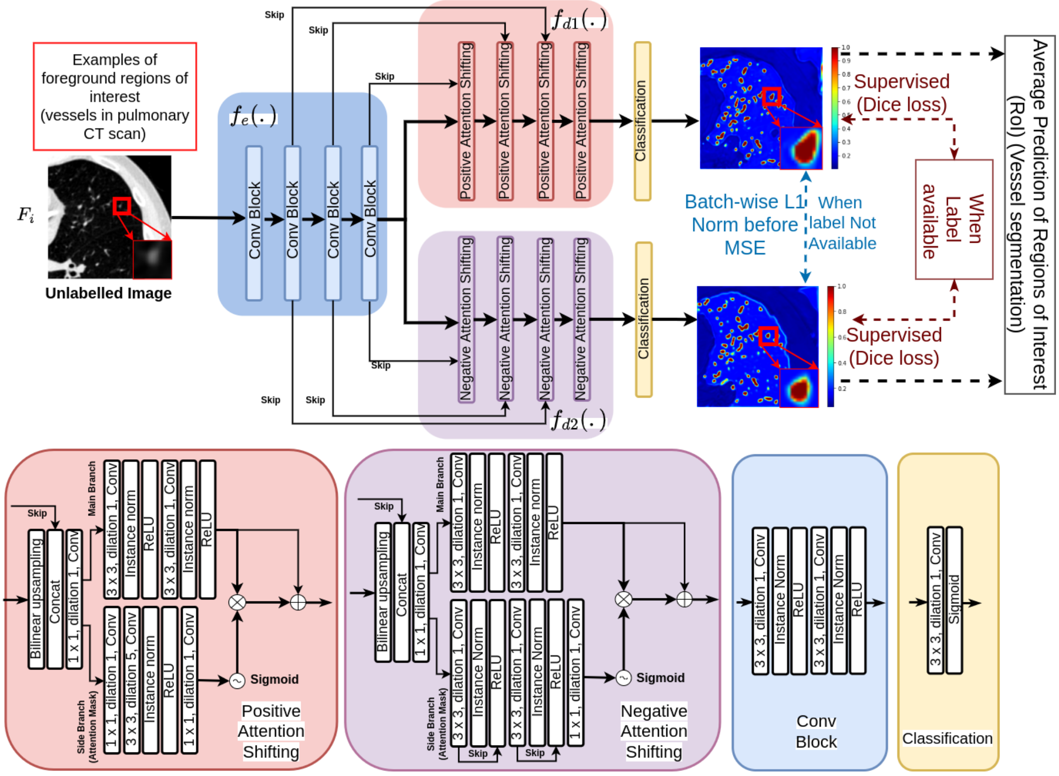
Examples of foreground regions of interest (vessels in pulmonary CT scan)
F_i
Unlabelled Image
$f_e(.)$
Conv Block
Skip
$f_{d1}(.)$
Positive Attention Shifting
$f_{d2}(.)$
Negative Attention Shifting
Classification
Supervised (Dice loss)
Batch-wise L1 Norm before MSE
When label Not Available
When Label available
Average Prediction of Regions of Interest (RoI) (Vessel segmentation)
Main Branch
Side Branch (Attention Mask)
Bilinear upsampling
Concat
1 x 1, dilation 1, Conv
3 x 3, dilation 1, Conv
3 x 3, dilation 5, Conv
Instance norm
ReLU
Sigmoid
Positive Attention Shifting
Negative Attention Shifting
Conv Block
Classification

Figure 4.4: MisMatch (U-net based): decoder f_{d1} leads to dilated high confidence detection of foreground and decoder f_{d2} leads to eroded high confidence detection of foreground. The final prediction is the average between outputs of f_{d1} and f_{d2}. Any other encoder-decoder segmentation network could be used.

Overview of MisMatch In this paper, we learn to realistically morph the foreground features by controlling the ERF for consistency regularisation. In order to create a paired predictions with different confidences for consistency regularisation, our strategy is to dilate the foreground features and erode the foreground features, we also compare our strategy with other possible strategies in an ablation study in later section 5.7. As introduced in the last section, the prediction confidence can be affected by the ERF while the ERF is decided by the network topology. More specifically, we use the dilated convolutional layer to raise the ERF on one hand to dilate the foreground features, and we use skip-connections to decrease the ERF on the other hand to erode the features of foreground. However, we do not know how much confidence should be changed at each pixel. To address this, we introduce soft attention mechanism to learn the magnitude of the confidence change for each pixel. Now we introduce how we achieve this in the next section.

Differences between proposed methods and classical morphological operations We also would like to highlight the difference between our approach at feature space and the classical morphological operations at image space. Traditional morphological operations simply remove/add boundary pixels using local neighbouring information which is not differentiable, in contrast, our approach is differentiable and can be fully integrated in neural networks.

4.4.2 Architecture of Mismatch

As shown in Fig.4.4, MisMatch is a framework which can be integrated into any encoder-decoder based segmentation architecture. In this section, we use 2D U-net [62] due to its popularity in medical imaging, although later we also have an experiment using a MisMatch based on a 3D V-net. Our U-net based MisMatch (**Fig 4.4**) has two components, an encoder (f_e) and a two-head decoder (f_{d1} and f_{d2}). The first decoder (f_{d1}) comprises of a series of *Positive Attention Shifting Blocks*, which shifts more attention towards the foreground area, resulting in dilating high-confidence predictions on the foreground. The second decoder (f_{d2}) containing a series of *Negative Attention Shifting Blocks*, shifts less attention towards the foreground, resulting in eroding high-confidence predictions on the foreground.

4.4.3 Positive Attention Shifting Block

The positive Attention Shifting Block aims at increasing the ERF of the foreground, therefore dilating the foreground features. In a standard U-net, a block ($f(.)$) in the decoder comprises two consecutive convolutional layers with kernel size (K) 3 followed by ReLU and normalisation layers. If the input of $f(.)$ is x and the output of $f(.)$ is $f(x)$, to increase the ERF of $f(x)$, we would aim to generate an attention mask with a larger ERF than the ERF of $f(x)$. To do so, we add a parallel side branch $f'(.)$ next to the main branch $f(.)$. The side branch intakes x but outputs $f'(x)$ with a larger ERF. We apply Sigmoid on the output of the side branch as an attention mask to increase the confidence of $f(x)$. The new block containing both

$f(.)$ and $f'(.)$ is our proposed Positive Attention Shifting Block (PASB). The side branch of the PASB is a dilated convolutional layer with dilation rate 5.

4.4.3.1 ERF size in Positive Attention Shifting Block

Given the size of ERF of n^{th} layer as, $\sqrt{n}$ [122], which is the input x, as output from the previous layer. The ERF of $f(x)$ is $ERF_{f(x)} = K\sqrt{n+2}$. To make sure the ERF of $f'(x)$ is larger than $K\sqrt{n+2}$:

$$\frac{ERF_{f'(x)}}{ERF_{f(x)}} = \frac{K'}{K}\sqrt{\frac{1}{1+\frac{1}{n+1}}} > \lim_{n\to+0}\frac{K'}{K}\sqrt{0.5} > 1 \tag{4.1}$$

From Eq4.1, we find $K' > \frac{1}{\sqrt{0.5}}K \approx 1.5K$. We double the condition as our design choice, then K' is 9 when $K = 3$. However, the large kernel sizes significantly increase model complexity. To avoid this, we use a dilated convolutional layer to achieve K' at 9, which requires a dilation rate 5. As the side branch has a larger ERF than the main branch, it can raise the confidence on the foreground of the main branch. Previous work [123, 124] has reported similar uses of a dilated convolutional layer to increase the ERF for other applications, without explaining the rationale for their use. See visual evidence in Fig 4.4(q) and (r).

4.4.4 Negative Attention Shifting Block

The negative Attention Shifting Block aims at decreasing the ERF on the foreground, therefore eroding the foreground features. Following PASB, we design the Negative Attention Shifting Block (NASB) again as two parallel branches. In NASB, we aim to shrink the ERF of the $f(x)$ in order to produce a smaller ERF than the one from the main branch. In the side branch in NASB, we use the same architecture as the main branch, but with skip-connections as skip-connections restrict the growth of the ERF with increasing depth [122].

4.4.4.1 ERF size in Negative Attention Shifting Block

Neural networks with residual connections are equivalent to an ensemble of networks with short paths where each path follows a binomial distribution [125]. If we define p as the probability of the model going through a convolutional layer and

$1-p$ as the probability of the model skipping the layer, then each short path has a portion of $\binom{N}{k}p^k(1-p)^{n-k}$, contributing to the final ERF. If we assume p is 0.5, the ERF of the side branch is guaranteed to be smaller than the ERF of the main branch, see Eq.4.2.

$$\begin{aligned} \frac{ERF_{f'(x)}}{ERF_{f(x)}} &= 0.25\sqrt{\frac{1}{1+\frac{2}{n}}}+0.5\sqrt{\frac{1}{1+\frac{1}{n+1}}}+0.25 \\ &< \lim_{n\to+\infty} 0.25+0.5+0.25 = 1 \end{aligned} \tag{4.2}$$

As the side branch has a smaller ERF than the main branch, it can reduce the confidence on the foreground of the main branch. See visual evidence in Fig 4.4(u) and (v).

4.4.5 Loss Functions

For experiments on BRATS 2018 and CARVE 2014, We use a streaming training setting to avoid over-fitting on limited labelled data so the model doesn't repeatedly see the labelled data during each epoch. When a label is available, we apply a standard Dice loss [126] between the output of each decoder and the label. When a label is not available, we apply a mean squared error loss between the outputs of the two decoders. This consistency regularisation is weighted by hyper-parameter α. For experiments on LA 2018, we train simultaneously on labelled and unlabelled images by combine consistency regularisation loss with Dice loss.

4.5 Experiments

We perform a few sets of experiments: 1) comparisons with baselines including supervised learning and state-of-the-art SSLs [2, 1, 82, 4] using either data or feature augmentation; 2) investigation of the impact of the amount of labelled data and unlabelled data on MisMatch performance; 3) ablation study of the decoder architectures; 4) ablation study on the hyper-parameter such as α

4.5.1 Data sets & Pre-processing

CARVE 2014 The Classification of pulmonary arteries and veins (CARVE) dataset [127] has 10 fully annotated non-contrast low-dose thoracic CT scans. Each case has between 399 and 498 images, acquired at various spatial resolutions between (282 x 426) to (302 x 474). 10-fold cross-validation on the 10 labelled cases is performed. In each fold, we split cases as: 1 for labelled training data, 3 for unlabelled training data, 1 for validation and 5 for testing. We only use slices containing more than 100 foreground pixels. We prepare datasets with differing amounts of labelled slices: 5, 10, 30, 50, 100. We crop 176 × 176 patches from four corners of each slice. Full label training uses 4 training cases. Normalisation was performed case wise.

BRATS 2018 BRATS 2018 [128] has 210 high-grade glioma and 76 low-grade glioma MRI cases, each case containing 155 slices. We focus on binary segmentation of whole tumours in high grade cases. We randomly select 1 case for labelled training, 2 cases for validation and 40 cases for testing. We centre crop slices at 176 × 176. For labelled training data, we extract the first 20 slices containing tumours with areas of more than 5 pixels. To see the impact of the amount of unlabelled training data, we use 3100, 4650 and 6200 slices respectively. Case-wise normalisation was performed and all modalities were concatenated. We train each model 3 times and take the average.

LA 2018 Atrial Segmentation Challenge Data set [129] has 100 volumes of 3D gadolimium-enhanced MR scans with corresponding left atrium segmentation masks. Each scan is isotropic with resolution at 0.625 x 0.625 x 0.625 mm^3. We follow [130] and split 100 scans into 80 for training and 20 for testing. We also directly use the pre-processing from [130] to normalise the centre crop each scan.

Task 01 Brain Tumour Task01 Brain Tumour from Medical Segmentation Decathlon consortium [131] is based on BRATS 2017 with different naming format from BRATS 2018. Each case in The Task01 Brain Tumour has 155 slices with 240 x 240 spatial dimension. We merge all of the tumour classes into one tumour class for simplicity. We do not apply centre cropping in the pre-processing here. In the training, we randomly crop volumes on the fly with size of 96 x 96 x 96. We separate the original training cases as labelled training data and testing data. We use the original testing cases as unlabelled data. For the labelled training data, we use 8 cases with index number from 1 to 8. We have 476 cases for testing and 266 cases for unlabelled training data. We apply normalisation with statistics of intensities across the whole training data set. We keep all of the MRI modalities as 4 channel input.

Table 4.1: MisMatch (MM) vs Baselines on CARVE. Metric is Intersection over Union (IoU).

	Supervised		**Semi-Supervised**					
Labelled **Slices**	**Sup1** [62](2015)	**Sup2** Ours(2021)	**MTA** [82](2019)	**MT** [1](2017)	**FM** [2](2020)	**CCT** [4](2020)	**Morph** 2021	**MM** **Ours(2021)**
5	48.32±4.97	50.75±2.0	54.91±1.82	56.56±2.38	49.30±1.81	52.54±1.74	52.93±2.19	**60.25±3.77**
10	53.38±2.83	55.55±4.42	57.78±3.66	57.99±2.57	51.53±3.72	55.25±2.52	57.08±2.96	**60.04±3.64**
30	52.09±1.41	53.98±4.42	60.78±4.63	60.46±3.74	55.16±5.93	60.81±4.09	60.19±4.97	**63.59±4.46**
50	60.69±2.51	64.79±3.46	68.11±3.39	67.21±3.05	62.91±6.99	65.06±3.42	64.88±3.25	**69.39±3.74**
100	68.74±1.84	73.1±1.51	72.48±1.61	71.48±1.57	72.58±1.84	72.07±1.75	72.11±1.88	**74.83±1.52**
Param. (M)	1.8	2.7	2.1	1.88	1.88	1.88	2.54	2.7
Infer.Time(s)	4.1e-3	1.8e-1	7.2e-3	4.3e-3	4.5e-3	1.5e-1	8e-3	1.8e-1

Table 4.2: MisMatch (MM) vs Baselines on BRATS. Metric is Intersection over Union (IoU).

	Supervised		**Semi-Supervised**					
Unlabelled **Slices**	**Sup1** [62](2015)	**Sup2** Ours(2021)	**MTA** [82](2019)	**MT** [1](2017)	**FM** [2](2020)	**CCT** [4](2020)	**Morph** 2021	**MM** **Ours(2021)**
3100	53.74±10.19	55.76±11.03	50.53±8.76	55.29±10.21	57.92±12.35	56.61±11.7	53.88±9.99	**58.94±11.41**
4650	53.74±10.19	55.76±11.03	47.36±6.65	58.32±12.07	54.29±9.69	56.94±10.93	55.82±11.03	**60.74±12.96**
6200	53.74±10.19	55.76±11.03	50.11±8.00	56.92±12.20	56.78±11.39	57.37±11.74	54.5±9.75	**58.81±12.18**

4.5.2 Implementation

We use Adam optimiser [58]. Hyper-parameters are: $\alpha = 0.002$, batch size 1 (GPU memory: 2G), learning rate 2e-5, 50 epochs. Each complete training on CARVE takes about 3.8 hours. The final output is the average of the outputs of the two decoders. In testing, we take an average of models saved over the last 10 epochs across experiments. Our code is implemented using Pytorch 1.0 [132].

4.5.3 Baselines

In the current study the backbone is a 2D U-net [62] with 24 channels in the first encoder. To ensure a fair comparison we use the same U-net as the backbone across all baselines. The first baseline utilises supervised training on the backbone, is

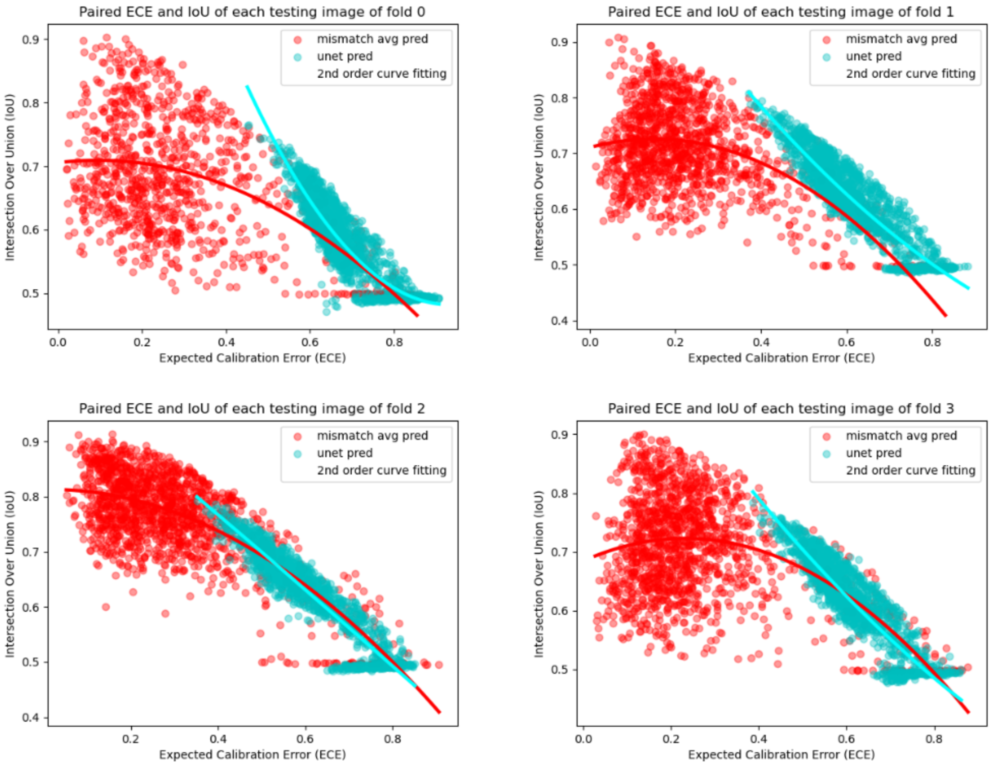

Figure 4.5: Expected calibration error [5] against accuracy in 10-fold cross-validation experiments on 50 labelled slices with CARVE. Y-axis: IoU. X-axis: ECE. Each calibration error is calculated from the gap between the confidence and accuracy for each testing image. Each data point in this figure is one testing image. The fitted 2nd order trends of our MisMatch are flatter than U-net, meaning MisMatch is more robust against the calibration error.

trained with labelled data, augmented with flipping and Gaussian noise and is denoted as “Sup1”. To investigate how unlabelled data improves performance, our second baseline “Sup2” utilises supervised training on MisMatch, with the same augmentation. Because MisMatch uses consistency regularisation, we focus on comparisons with five consistency regularisation SSLs: 1) “mean-teacher” (MT) [1], with Gaussian noise, which has inspired most of the current state-of-the-art SSL methods; 2) the current state-of-the-art model called “FixMatch” (FM) [2]. To adapt FixMatch for a segmentation task, we use Gaussian noise as weak augmentation and “RandomAug” [133] without shearing for strong augmentation. We do not use shearing for augmentation because it impairs spatial correspondences of pixels of paired dense outputs; 3) a state-of-the-art model with multi-head decoder

[4] for segmentation (CCT), with random feature augmentation in each decoder [4]. This baseline is also similar to models recently developed [3, 76]; 4) a further recent model in medical imaging [82] using image reconstruction as an extra regularisation (MTA), augmented with Gaussian noise; 5) a U-net with two standard decoders, where we respectively apply erosion and dilation on the features in each decoder, augmented with Gaussian noise (Morph)"; 6) an uncertainty aware mean-teacher based SSL segmentation model [130]. Our MisMatch model has been trained without any augmentation.

4.6 Segmentation Results

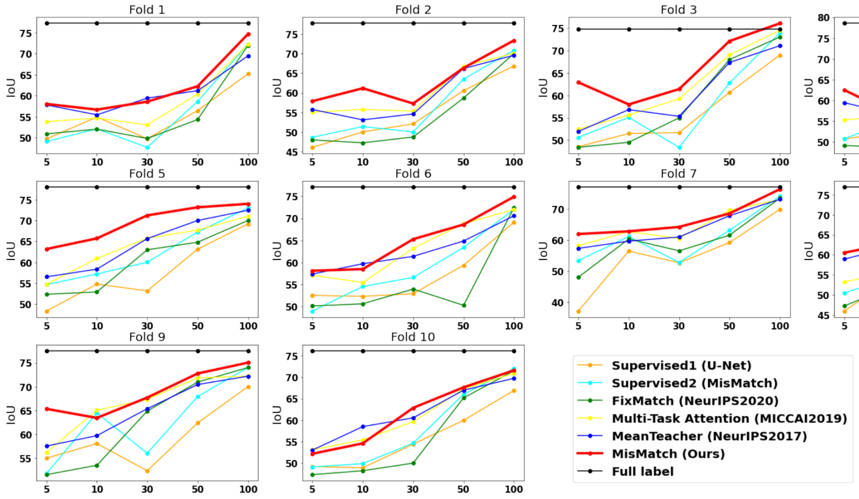

Figure 4.6: Full results of 10 fold cross-validation on CARVE. X-axis: number of labelled slices. Y-axis: IoU

MisMatch consistently and substantially outperforms supervised baselines, the improvement is especially obvious in low data regime. For example, on 5 labelled slices with CARVE, MisMatch achieves 24% improvement over Sup1. MisMatch consistently outperforms previous SSL methods [2, 1, 82, 4] in Table 5.2, across different data sets. Particularly, there exists statistical difference between Mismatch and other baselines when 6.25% labels (100 slices comparing to 1600 slices of full label) are used on CARVE (Table 4.3). Qualitatively, we observed in Fig 4.8 that, the main performance boost of MisMatch comes from the reduction of false positive detection and the increase of true positive detection.

Interestingly, we found that Sup2 (supervised training on MisMatch without unlabelled data) is a very competitive baseline comparing to previous semi-supervised methods. This might imply that MisMatch can potentially help with the supervised learning as well.

We also found data diversity of training data highly affects the testing performance (Fig 4.6) in cross-validation experiments. For example, in fold 3, 7 and 8 on CARVE, MisMatch outperforms or performs on-par with the full label training, whereas in the rest folds, MisMatch performs marginally inferior to the full label training. Additionally, more labelled training data consistently produces a higher mean IoU and lower standard deviation (Table 5.3). Lastly, we noticed more unlabelled training data can help with generalisation, until it dominates training and impedes performance (Table 5.3).

We further verify that consistency regularisation on feature perturbations is better than consistency regularisation on input perturbations by comparing MisMatch against UA-MT [130] which is an representative example of the methods using input perturbations. We compare MisMatch against UA-MT on two 3D datasets left atrium and whole tumour areas (see section 4.5.1). On the segmentation on left atrium, our method not just outperform UA-MT but also converges faster, as illustrated in Fig.4.9.

During testing of trained models on the whole tumour segmentation from the Task01 Brain Tumour data set[131], we noticed one emerging property of our model that the our model achieves better performance when it is tested on volumes larger than the size of the training volumes (see Table 4.7 and Table 4.8). Also if the testing size is smaller than the training size, the performance becomes worse (see Table 4.7 and Table 4.9).

Table 4.3: P-value between MM and baselines. Non-parametric Mann-Whitney U-Test. 100 labelled slices of CARVE.

Sup1	**Sup2**	**MTA**	**MT**	**FM**	**CCT**	**Morph**
9.13e-5	1.55e-2	4.5e-3	4.3e-4	1.05e-2	1.8e-3	2.2e-3

4.6.1 Ablation Studies

We performed ablation studies on the architecture of the decoders of MisMatch with cross-validation on 5 labelled slices of CARVE: 1) “MM-a”, a two-headed U-net with standard convolutional blocks in decoders, the prediction confidences of these two decoders can be seen as both normal confidence, however, they are essentially slightly different because of random initialisation, we denote the decoder of U-net as f_{d0}; 2) “MM-b”, a standard decoder of U-net and a negative attention shifting decoder f_{d2}, this one can be seen as between normal confidence and less confidence; 3) “MM-c”, a standard decoder of U-net and a positive attention shifting decoder f_{d1}, this one can be seen as between normal confidence and higher confidence; 4) “MM”, f_{d1} and f_{d2} (Ours). As shown in Fig 4.7, our MisMatch (”MM”) outperforms other combinations in 8 out of 10 experiments and it performs on par with the others in the rest 2 experiments. Among the results when MisMatch outperforms, MisMatch outperforms MM-a by 2%-14%; outperforms MM-b by 3%-18%; outperforms MM-c by 4%-22%. We also tested α at 0, 0.0005, 0.001, 0.002, 0.004 with the same experimental setting. The optimal α appears at 0.002 in Table 4.4. We also found that gradient cutting helps to improve segmentation performance too, see Table 4.6. In terms of network topology, as shown in Table 4.5 , it seems that larger dilation is not always beneficial.

Table 4.4: Ablation studies on alpha value using CARVE with 5 labelled slices.

alpha	**0.0**	**0.0005**	**0.001**	**0.002**	**0.004**
IoU	50.75	59.16	59.45	60.25	58.89

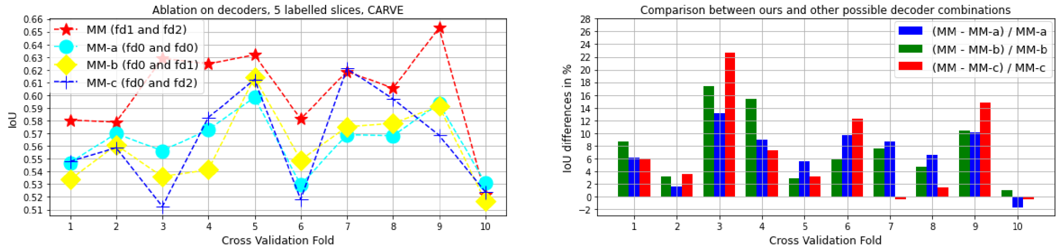

Figure 4.7: Ablation studies on decoder architectures, cross-validation on 5 labelled slices with CARVE. MM is ours.

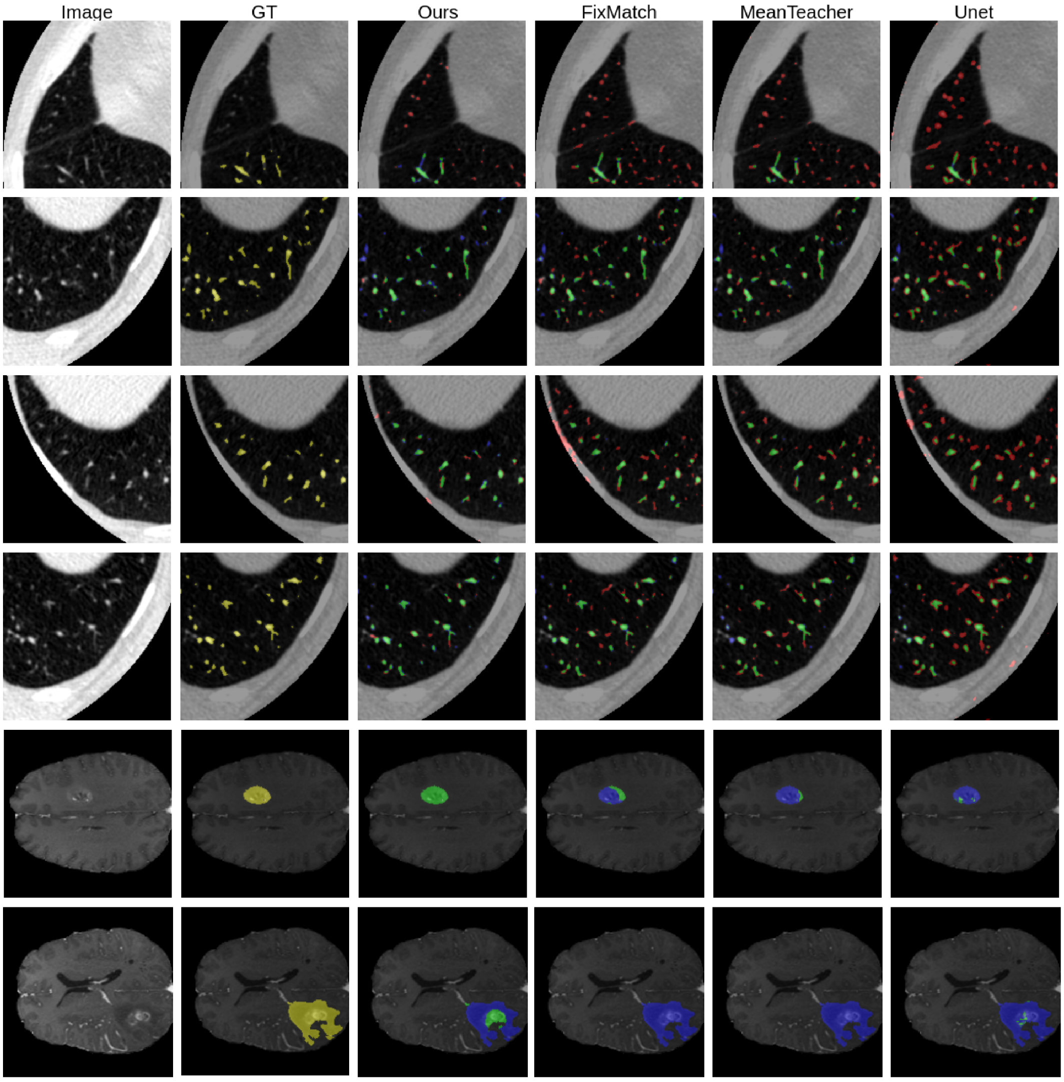

Figure 4.8: Visual results. Yellow: ground truth. Red: False Positive. Green: True Positives. Blue: False Negatives. Row 1-4: CARVE. Row 5-6: BRATS

Table 4.5: Ablation studies on dilation rate in 3D V-net based MisMatch using LA 2018 with 2 labelled cases, α as 1 and cutting gradients, network width 8. Metric is Dice score.

Iteration	**2000**	**3000**	**4000**	**5000**
Dilation 6	0.6571	0.6621	0.6699	0.6561
Dilation 9	0.7363	0.7283	0.7180	0.6561
Dilation 12	0.6980	0.6957	0.6889	0.6561

Table 4.6: Ablation studies on stopping gradients in 3D V-net based MisMatch using LA 2018 with 2 labelled cases, α as 1, network width 8. Metric is Dice score.

Iteration	**2000**	**3000**	**4000**	**5000**
Stop gradient	0.6896	0.7148	0.7090	0.7057
Gradient	0.6717	0.6944	0.6952	0.6837

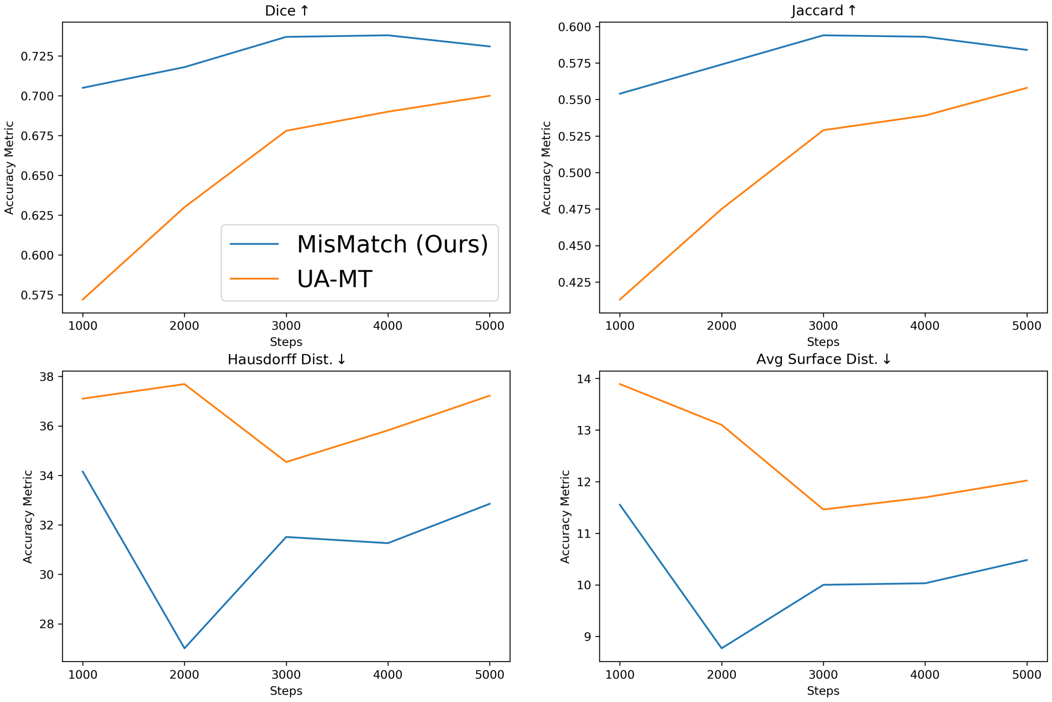

Figure 4.9: Results on LA 2018 between UA-MT and MisMatch with 2 labelled cases, lr 0.01, batch 4, consistency 1 and network width 8. This further confirms that consistency regularisation on feature perturbations is more effective than consistency on input perturbations.

Table 4.7: Testing Results on 3D segmenting the whole tumour from Task 01 Brain Tumour from Medical Segmentation Decathlon. Training with learning rate 0.001 and 3500 epochs. Testing on $96 \times 96 \times 96$ cubes. Jac: Jaccard. HD: Hausdorff Distance. ASD: Average Surface Distance.

Metrics	**Dice (↑)**	**Jac (↑)**	**HD (↓)**	**ASD (↓)**
UA-MT	0.5454	0.3864	55.25	22.74
MisMatch (Ours)	0.57	0.4197	49.07	22.86

4.7 Visualisation of the effectiveness of Learnt Attention Masks

We visualise the probabilistic certainty of foreground feature maps before and after attention, attention weights and how much the certainty are changed in Fig4.10 on

Table 4.8: Testing Results on 3D segmenting the whole tumour from Task 01 Brain Tumour from Medical Segmentation Decathlon. Training with learning rate 0.001 and 3500 epochs. Testing on 48 × 48 × 96 cubes. Jac: Jaccard. HD: Hausdorff Distance. ASD: Average Surface Distance.

Metrics	Dice (↑)	Jac (↑)	HD (↓)	ASD (↓)
UA-MT	0.2926	0.1769	72.66	31.98
MisMatch (Ours)	0.3133	0.1944	85.35	39.27

Table 4.9: Testing Results on 3D segmenting the whole tumour from Task 01 Brain Tumour from Medical Segmentation Decathlon. Training with learning rate 0.001 and 3500 epochs. Testing on 128 × 128 × 96 cubes. Jac: Jaccard. HD: Hausdorff Distance. ASD: Average Surface Distance.

Metrics	Dice (↑)	Jac (↑)	HD (↓)	ASD (↓)
UA-MT	0.5945	0.4390	54.88	22.15
MisMatch (Ours)	0.6086	0.4650	47.66	23.58

CARVE. We focus on zoomed-in area of one vessel which is one region-of-interest of the foreground. As shown in (c) and (e), the certainty outputs between the two decoders are different, the one from the positive attention decoder has more detected high certainty areas on the top of the anatomy of the interest. As illustrated in (j) and (n), the attention weights in the two decoders are drastically different from each other. More specifically, the attention weights in the negative attention decoder have relatively low values around the edges, as shown in green and blue colours, on the contrary, the attention weights in the positive attention decoder have high values in most of the regions of the interest.

Another evidence supporting the effectiveness of attention blocks are the changes of the certainty as shown in (r) and (v). After positive attention weights are applied on (g), it is clear to see in (r) that the surrounding areas of the originally detected contours are now also detected as regions of the interest. Besides, in (v), we observe expected negative changes of the certainty around edges caused by the negative attention shifting.

The histograms of the feature maps also support the effectiveness of our learnt attention masks. Between the histograms in (j) and (m), for the high certainty interval between 0.9 and 1.0, the negative attention block has more high uncertainty pixels than the positive attention block. This is because the negative attention block

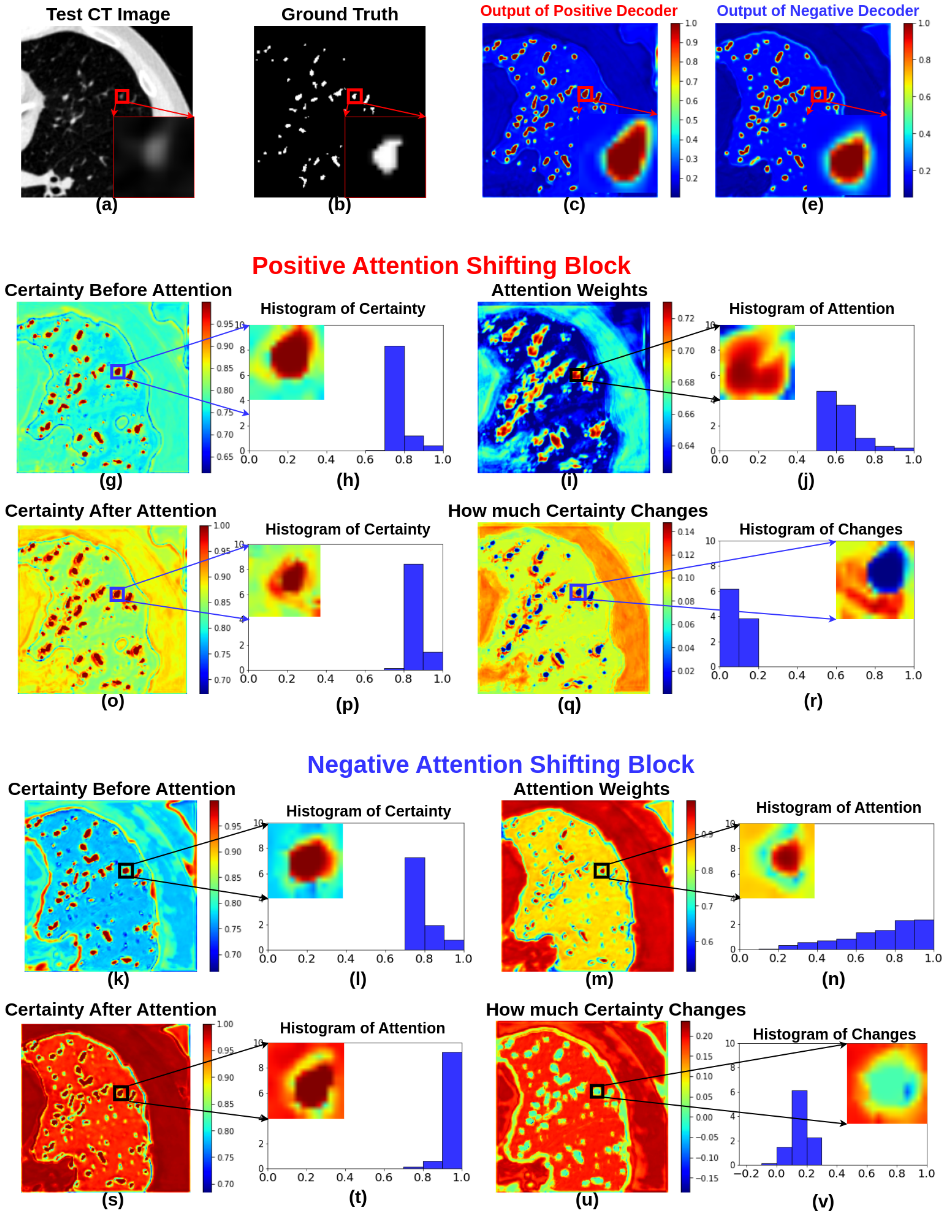

Figure 4.10: Visulisation of predicted certainty of the foreground in the last positive attention shifting decoder and the last negative attention shifting decoder. We focus on the zoomed-in regions on the foreground area containing one vessel.

decreases certainty on foreground, thereby ending up with increasing certainty on background, where background class is the majority class naturally containing more pixels than the foreground class.

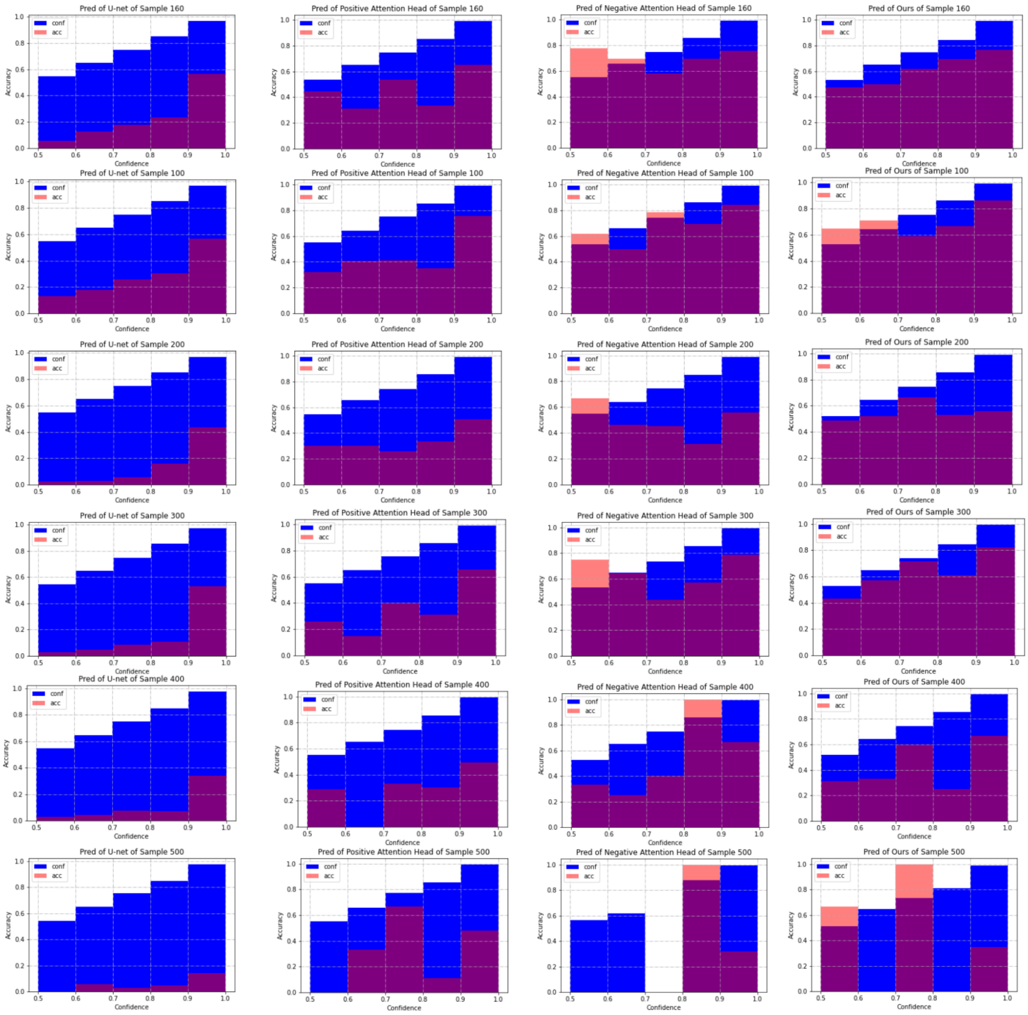

Figure 4.11: Reliability diagrams [5] from experiments on 50 labelled slices with CARVE. Blue: Confidence. Red: Accuracy. Each row is on one testing image. X-axis: bins of prediction confidences. Y-axis: accuracy. Column 1: U-net. Column 2: outputs of positive attention decoders. Column 3: outputs of negative attention decoders. Column 4: average outputs of the two decoders. The smaller the gap between the accuracy and the confidence, the better the network is calibrated.

4.8 Confidence and Calibration of Mismatch

Expected Calibration Error To qualitatively study the confidence of MisMatch, we adapt two mostly used metrics in the community, which are Reliability Diagrams and Expected Calibration Error (ECE) [5]. Following [134], we first prepare M interval bins of predictions. In our binary setting to classify the foreground, we use 5 intervals between 0.5 to 1. Say B_m is the subset of all pixels whose prediction confidence is in interval I_m. We define accuracy as how many pixels are correctly classified in each interval. The accuracy of B_m is formally:

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \tag{4.3}$$

Where $\hat{y}_i$ is the predicted label and y_i is the ground truth label at pixel i in B_m. The average confidence within B_m is defined with the use of $\hat{p}_i$ which is the raw probability output of the network at each pixel:

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \tag{4.4}$$

Ideally, we would like to see $conf(B_m) = acc(B_m)$, which means the network is perfectly calibrated and the predictions are completely trustworthy. To assess how convincing the prediction confidences are, we calculate the gap between confidence and accuracy as Expected Calibration Error (ECE):

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{4.5}$$

MisMatch is well-calibrated and effectively learns to change prediction confidence As shown in Fig4.11, both positive attention shifting decoder and negative attention shifting decoder are better calibrated than the plain U-net. Especially, positive attention shifting decoder produces over-confident predictions. Meanwhile, negative attention shifting decoder produces under-confident predictions for a few confidence intervals. This verifies again that MisMatch can effectively learn to differently change the prediction confidences of the same testing images.

Robustness of MisMatch Against Calibration Errors As shown in the scatter plot (Fig4.5) of paired IoU and corresponding Expected Calibration Error (ECE) of all of the testing images in cross-validation experiments on 50 labelled slices of CARVE, higher calibration errors correlate positively with low segmentation accuracy. In general, MisMatch has predictions with less calibration errors and higher IoU values. As shown in the 2nd order regression curves for each trend, MisMatch appears to be more robust against calibration error, as the fitted curve of U-net has a much more steep slope than MisMatch. In other words, with the increase of calibration error, MisMatch suffers less performance drops.

4.9 Conclusion

We propose MisMatch, an augmentation-free SSL, to overcome the limitations associated with consistency-driven SSL in medical image segmentation. In lung vessel segmentation tasks, the acquisition of labels can be prohibitively time-consuming. For example each case may take 1.5 hours of manual refinement with semi-automatic segmentation[127]. Longer timeframes may be required for cases with severe disease. MisMatch requires 100 slices of one case for training whereas the fully labelled dataset comprises 1600 slices across 4 cases. MisMatch when trained on just 10% of labels achieves a similar performance (IoU: 75%) to models that are trained with all available labels (IoU: 77%).

Chapter 5

Expectation-Maximization Pseudo Labelling for Segmentation with Limited Annotations

This chapter provides a new perspective of pseudo labelling through the lens of Bayesian statistics. This chapter is based on a publication which was shortlisted for Young Scientist Award (top 0.8% among submissions) at the 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2022. The current form of this chapter is under review at Medical Image Analysis Special Issue on MICCAI 2022. I conceived the idea, implemented the code, performed the experiments and wrote the draft of the manuscript; my colleagues Yukun, Chen and Yipeng provided feedback on experiments designs and the notations; all of the co-authors contributed to the writing of the manuscript.

5.1 Abstract

We study pseudo labelling and its generalisation for semi-supervised segmentation of medical images. Pseudo labelling has achieved great empirical successes in semi-supervised learning, by utilising raw inferences on unlabelled data as pseudo labels for self-training. In our paper, we build a connection between pseudo labelling and the Expectation Maximization algorithm which partially explains its empirical successes. We thereby realise that the original pseudo labelling is an empirical estima-

tion of its underlying full formulation. Following this insight, we demonstrate the full generalisation of pseudo labels under Bayes' principle, called Bayesian Pseudo Labels. We then provide a variational approach to learn to approximate Bayesian Pseudo Labels, by learning a threshold to select good quality pseudo labels. In the rest of the paper, we demonstrate the applications of Pseudo Labelling and its generalisation Bayesian Psuedo Labelling in semi-supervised segmentation of medical images on: 1) 3D binary segmentation of lung vessels from CT volumes; 2) 2D multi class segmentation of brain tumours from MRI volumes; 3) 3D binary segmentation of brain tumours from MRI volumes. We also show that pseudo labels can enhance the robustness of the learnt representations.

5.2 Introduction

5.2.1 Label scarcity: one major bottleneck of deep learning in medical imaging

Medical image segmentation is to accurately label pixels of interest in medical images, as a foundational step towards a broad range of downstream tasks such as computer-aided diagnosis, real-time surgical navigation, drug discovery and so on. Recent years have seen rises of deep learning enabled medical image segmentation methods for more accurate and faster segmentation results. However, there is no free lunch. According to the scaling laws [135], given fixed computational resources and model sizes, the performances of deep learning models depend on the amount of paired data and labels. Unfortunately, it is notoriously difficult to acquire a large amount of annotations for segmentation of medical images because of the high costs of money and time. Unlike natural images which are normally 2D, medical images are usually volumetric with high resolution, making the pixel-level annotation process exponentially more labour intensive than standard image segmentation tasks in computer vision. In addition, the annotation process of medical images require highly skilled medical experts which are extra costly.

To address the unavoidable label scarcity issue in medical image segmentation, semi-supervised learning has been introduced to jointly train on both labelled

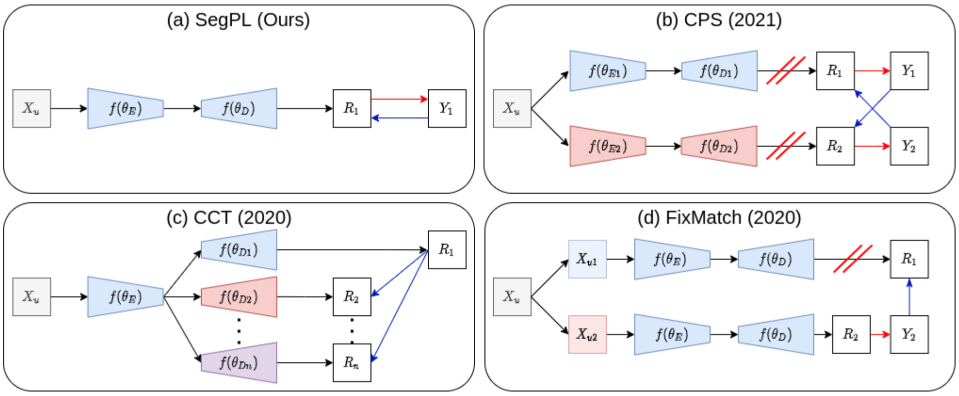

Figure 5.1: Comparison between pseudo labelling approach (SegPL) and the other approaches. Pseudo labelling approaches enjoy simplicity in implementation including corresponding benefits such as scalibility and robustness.

and unlabelled data to bootstrap the model performance. The unlabelled images normally have a much larger quantity than the labelled ones, but they are normally ignored in supervised learning. Semi-supervised learning is attractive because it improves the model by using the existing resource of unlabelled images, avoiding investments in label acquisitions. Apart from semi-supervised learning for addressing label scarcity, other approaches have also been invented such as outsourcing data labelling [136] combined with federated learning, which cannot really avoid labelling costs. However, semi-supervised learning still remains attractive as it has a good trade-off balance between the cost and the improvements of the models. The current paper focuses on semi-supervised learning.

5.2.2 Brief review of semi-supervised learning

Most of the semi-supervised learning methods aim at minimising the entropy of predictions of unlabelled data in order to make "firm" predictions. Entropy minimisation has a long history in representation learning. One of the earliest form of entropy minimisation was derived from the mutual information between the input and output in unsupervised learning [120]. Entropy regularisation gained tremendous popularity in semi-supervised image classification after the authors in [66] proposed to minimise the entropy on unlabelled data as a strong regularisation to drive the model to learn a good decision boundary. Since then, entropy minimisation has been evolving from its original explicit form to varying implicit forms.

Among different implicit forms of entropy minimisation, consistency regularisation is the most common one and it is behind most of the recent state-of-the-art methods in semi-supervised classification and segmentation [2, 69, 75, 3, 137]. There are mainly two types of consistency regularisation, one is soft consistency regularisation which applies distance based loss directly on the raw outputs or the probabilities of predictions, the other type is hard consistency regularisation which transforms raw outputs to pseudo label to supervise the raw outputs or the probabilities of predictions. Both types of the consistency regularisation [1, 2, 69, 138, 137] enforce the deep learning models to make predictions which are invariant to the perturbations at the input level or the feature level [4, 3, 139, 137]. For methods with consistency regularisation on input level, a lot of them are derived from a classic method called Mean-Teacher [1]. The mean-teacher model has two models, the weight of the student model is the exponential moving averaging weights of the teacher model. The teacher model intakes a normal input while the student model intakes the same input but added with Gaussian noise, a mean square error is used for soft consistency regularisation between the outputs of the teacher and the student model. Another more advanced teacher-student model called FixMatch achieved state-of-the-art performance in semi-supervised classification [2]. In FixMatch [2], the model intakes two forward passes, one pass with weakly augmented (e.g. flipping) input and the other one pass with strongly augmented (e.g. shearing, random intensity) input. Then the output of the weakly augmented input will be used to generate a pseudo label as the ground truth for training the output of the strongly augmented input, the workflow of FixMatch is illustrated in (d) in Fig.5.1. Although FixMatch and its variants have achieved great successes in image classification, it was noticed that it was not straightforward to directly apply FixMatch-style methods in image segmentation, because the cluster assumption does not hold at pixel-level in dense prediction tasks [3]. To adopt consistency regularisation in segmentation, the authors in [4] have discovered that it is possible to apply perturbations on the features for instead of inputs before the consistency regularisation is applied. In [4], the authors directly apply augmentation techniques on the features

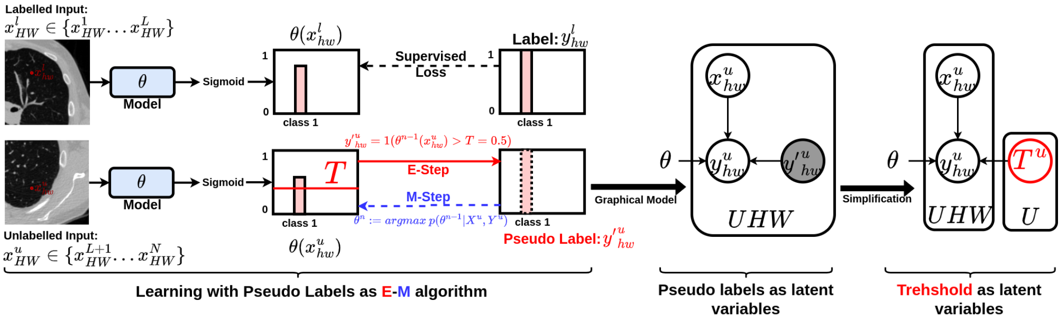

Figure 5.2: Pseudo-labelling process for binary segmentation. Pseudo-label y'_n is generated using unlabelled data x_u and model with parameters from last iteration θ. Therefore, pseudo-labelling can be seen as the E-step in Expecation-Maximization. The M-step updates θ using y'_n, y and data X. In our 1st implementation, namely SegPL, the threshold T is fixed for selecting the pseudo labels, which is the original pseudo labelling, as an empirical approximation of its true generalisation. In our 2nd implementation, namely SegPL-VI, the threshold T is dynamic and learnt via variational inference, which is an learnt approximation of its true generalisation.

of different decoders before the consistency regularisation (see (c) in Fig.5.1) for semi-supervised image segmentation. Apart from adding directly perturbations on the features directly, it is also feasible to add perturbations on the features through architectural modifications. For example, one can train two identical models but with different initialisation and apply consistency regularisation with pseudo labels on the two outputs [139] (Fig5.1 (b)) also for semi-supervised image segmentation. The aforementioned methods have also been tested and compared against our method in later section.5.7.

5.2.3 Motivations and contributions

It has come to our attention that most of the existing pseudo labelling papers are purely empirical without investigating the reason behind its empirical successes. Therefore we decided to revisit pseudo labelling and we noticed that pseudo labelling has a deep connection with the classical machine learning approach Expectation Maximization. Our second motivation is inspired by a recent paper in semi-supervised image classification, showing that it is entirely possible to achieve competitive results with smartly selected good quality pseudo labels [140]. In this paper, we build up our insight of pseudo labelling on the Expectation Maximization algorithm, we as well provide empirical investigations of pseudo labelling in semi-supervised medical image segmentation and its robustness. A shorter version of this paper has been published at MICCAI 2022 [101]. We summarize our contributions in the following bullet points.

- We interpret pseudo labelling within the framework of the Expectation Maximisation (EM) algorithm. As the EM algorithm is gauranteed to converge to local minimum. We therefore partially explain the empirical success of pseudo labelling.

- We provide a learning method to find the generalised pseudo labels using variational inference.

- We investigate the use of pseudo labelling in semi-supervised medical image segmentation and its characteristics such as robustness.

5.3 Related works

5.3.1 Pseudo Labelling

The original pseudo labelling [67] was proposed for semi-supervised multi-class image classification. In the original pseudo labelling, the pseudo labels are created as arg max outputs (same as running the inferences) on the predictions on the unlabelled data. The created pseudo labels are used to train the unlabelled data following standard supervised learning fashion. The original pseudo labelling actually creates the pseudo label online on-the-fly. It was also suggested to carefully use the pseudo labels that one should first warm-up the model with only supervised learning, then gradually ramp-up the weight of the pseudo supervision. Pseudo labelling has become popular in semi-supervised learning because it is computationally cheap but with good performance. It has been shown that semi-supervised learning can surpass its supervised learning counterpart [141] on ImageNet classification by using pseudo labelling on an internet-level unlabelled data. Another recent work breaks the trend of growing complexity of consistency regularisation methods and achieved competitive results with only pseudo labels[140] in semi-supervised image classification. In image segmentation, a pseudo labelling approach has also achieved impressive results [142] whereas the pseudo labels are refined with self-attention mechanism. Pseudo labelling also has its own potential issue which is called confirmation bias [143] that if wrong pseudo labels are used, noisy training will happen and the errors will be accumulated and potentially amplified. This potential issue inspired us and in later section, we will introduce a stochastic training of pseudo labels that learns to pick up correct pseudo labels.

5.3.2 Semi-supervised learning in medical image segmentation

Most of the existing methods for semi-supervised segmentation are stemmed from the methods in semi-supervised classification described in the previous section.5.2.2. For example, the mean-teacher based consistency regularisation methods have been popular in semi-supervised medical image segmentation [80, 71, 70, 73, 144, 145, 146, 147, 148, 149]. One of the earliest mean-teacher model in medical imaging was proposed by Yu [130], they improve on mean-teacher model by using uncertainty to generate a mask to apply consistency regularisation only on low-uncertainty areas. In addition to data level perturbations, feature level perturbations based consistency regularisation methods also popular. For instance, Luo [87] uses different initialisation of different decoders to achieve feature perturbations. Xu [137] uses differential morphological operations to add feature perturbations in the decoders before consistency regularisation. Pseudo labelling approaches have also been previously explored in medical image segmentation. Bai [150] used conditional random fields to remove the false positives of the pseudo labels. Wang [151] uses uncertainty to refine pseudo labels. Wu [152] combines pseudo labels with two headed network to form a cross pseudo supervision. Another recent work [114] uses a variational auto-encoder as student model and learns from pseudo labels which generated by deterministic teacher model.

5.4 Pseudo Labelling As Expectation-Maximization

In this section, we provide a new perspective of semi-supervised learning with pseudo labels as the Expectation-Maximization (EM) algorithm. We focus on a binary segmentation case as most of medical image segmentation tasks are binary ones on differentiating the foreground area from the background area. Multi-class segmentation can be trivially extended from the binary case with a multi-channel Sigmoid, by separetely treating each channel as a binary output, combining with the argmax operation before the final prediction.

5.4.1 Problem formulation

Given a set of N total available training images as $X = \{x_n \in R^{HW} : n \in (1,2,...,L,L+1,...,N)\}$, where $X_L = \{x_l \in R^{HW} : l \in (1,...,L)\}$ are L labelled images; $Y_L = \{y_l \in R^{HW} : l \in (1,...,L)\}$ are L labels for X_L; $X_U = \{x_u \in R^{HW} : u \in (L+1,...,N)\}$ is the rest of the U or $(N-L)$ unlabelled images. We have a segmentation network with parameters as θ and our final goal is to predict the labels $p(Y|X,\theta)$ of the whole data X with respect to θ.

5.4.2 Pseudo labels as latent variables

In order to find the optimal parameters of θ, the common approach is maximum likelihood estimation for maximising the likelihood of $P(X|\theta)$ with respect to θ, which contains two parts, namely the supervised learning part and unsupervised learning part. The supervised learning part is to find the following joint data density with known full information of the labels:

$$p(X_L, Y_L|\theta) \tag{5.1}$$

The unsupervised learning part is to find the underlying likelihood with the same parameters without full information of the data:

$$p(X_U|\theta) \tag{5.2}$$

Since labels are not observable for X_U, we can treat this as a missing data problem and introduce latent variables Y'_U. We therefore transform the above Eq. 5.2 to an estimation of the following marginal likelihood:

$$p(X_U|\theta) = \int p(X_U, Y'_U|\theta) dY'_U \tag{5.3}$$

We notice that the generation of pseudo-labels naturally poses a generative task. We also observe that the pseudo labels are only used as an intermediate step towards the final prediction of the labels. Thereby we propose to treat pseudo-labels as an implementation of the latent variables in the above Eq. 5.3. Eq. 5.3 also shows that it is not an easy task to train a model in semi-supervised fashion, because it is difficult to simultaneously estimate the optimal values of two different sets of variables. To address this difficult learning problem, we can decompose this problem by iteratively estimating the the latent variables Y'_U and the observed variable X_U. We now notice that this can be solved by a typical Expecation-Maximization (EM)[46] algorithm. By plugging the Jensen's inequality, one can iteratively refine the Evidence Lower Bound of the log likelihood of the data in Eq.5.3

5.4.3 E-M Pseudo Labelling

We now describe each component of the pseudo labelling in the sense of the EM algorithm in the following paragraphs.

E-step At the n^{th} iteration, the E-step estimates the posterior of the latent variable with the model (θ^{n-1}) from the last iteration ($n-1$). According to the cluster assumption that similar data points are supposed to have similar labels [153], the E-step runs the inference on unlabelled data and generate pseudo-labels according to its maximum predicted probability. In practice, in binary segmentation, the pseudo-labels for the foreground class 1 are picked using a fixed threshold value (T) between 0 and 1. Normally, this threshold is set up as 0.5. This binarization is actually equivalent to the plug-in principle [66], which is a common approach for estimating the posterior probability using an empirical estimation in statistics. Therefore, the pseudo-labelling itself is the E-step:

$$y_u^{hw'} = \mathbb{1}(\theta^{n-1}(x_u^{hw}) > T = 0.5) \tag{5.4}$$

The above equation Eq. 5.4 is pseudo-labelling at the pixel level. Where h and w are the index for the height and the index for the width of the pixel location respectively, for each unlabelled image x_u. $y_u^{hw'}$ is the pixel-wise pseudo label. More details of the connection between E-step and pseudo labelling is in the later section sec.5.4.4 on the convergence of pseudo labelling.

M-step At the M-step of iteration n, we will update the model parameters θ^{n-1} using the estimated latent variables (pseudo-labels Y_U') from the E-step. The images X are ignored for simplicity in the following expression (e.g. differing from the MICCAI version):

$$\theta^n := argmax_\theta p(\theta^{n_1} | \theta^{n-1}, Y_n') \tag{5.5}$$

The above Eq.5.5 is normally solved by setting the partial derivatives of the sum of the $p(Y_n')$ with respect to θ as zero, which can be calculated with modern automatic differentiation based deep learning toolbox such as Pytorch [132]. In practice, we solve Eq.5.5 via stochastic gradient descent. To use the stochastic gradient descent, we need to define an objective function and we use the common Dice loss ($f_{dice}(.)$) [126] as this is a segmentation task, where a is predicted probability and b is the label:

$$f_{dice}(a,b) = \frac{2 * a * b + \varepsilon}{a + b + \varepsilon} \tag{5.6}$$

Loss function of SegPL We weight the Eq.5.5 with a hyper-parameter α. For the whole data set including both unlabelled and labelled data, we can extend the Eq.5.5 and Eq.5.4 to a combination between the supervised learning part L_L and the unsupervised learning part L_U:

$$\mathscr{L}_{SegPL} = \underbrace{\alpha \frac{1}{N-L} \sum_{u=L+1}^{N} f_{dice}(\theta^{n-1}(x_u), \mathbb{1}(\theta^{n-1}(x_u) > T = 0.5))}_{\mathscr{L}_U} \\ + \underbrace{\frac{1}{L} \sum_{l=1}^{L} f_{dice}(\theta^{n-1}(x_l), y_l)}_{\mathscr{L}_L} \tag{5.7}$$

The above loss function 5.7 is the key component of our first proposed semi-supervised segmentation method, omitting pixels' locations for simplicity, which is referred as SegPL (Segmentation with Pseudo Labels) in the paper. $\mathscr{L}_L$ works to prevent the networks falling into trivial solutions, trivial solutions happen when networks constantly predict one single class for all of the pixels.

5.4.4 On the convergence of Pseudo Labelling from the perspective of EM

In this section, we explain how semi-supervised learning with pseudo-labelling will always converges from the perspective of EM. We first define a target function as $l(\theta)$, in our case, it would be the log likelihood of the data X. We also need to introduce a surrogate function $q(Y_U')$ which is any arbitrary distribution over the latent varaible Y_U'. We follow [46] and show the lower bound of the data likelihood in the form of the Free Energy of the generative latent variable model:

$$\begin{aligned} &\underbrace{l(\theta) log \int p(X_U, Y_U' | \theta) dY_U'}_{Eq.3} \\ &= log \int q(Y_U') \frac{p(X_U, Y_U' | \theta)}{q(Y_U')} dY_U' \\ &\geq \int q(Y_U') log \frac{p(X_U, Y_U' | \theta)}{q(Y_U')} dY_U' \\ &= \underbrace{\int q(Y_U') log p(X_U, Y_U' | \theta) dY_U'}_{Definition\, of\, Expectation} - \underbrace{\int q(Y_U') log q(Y_U') dY_U'}_{Entropy\, of\, surrogate function} \\ &\underbrace{\mathscr{F}(q(Y_U'), \theta)}_{Free\, Energy} \end{aligned} \tag{5.8}$$

We now show another decomposition of the data likelihood starting from the free energy:

$$
\begin{aligned}
&\mathscr{F}(q(Y_U'),\theta) \\
&= \int q(Y_U')log\frac{p(X_U,Y_U'|\theta)}{q(Y_U')}dY_U' \\
&= \int q(Y_U')log\frac{p(Y_U'|X_U,\theta)p(X_U|\theta)}{q(Y_U')}dY_U' \\
&= \int q(Y_U')logp(X_U|\theta)dY_U' + \int q(Y_U')log\frac{p(Y_U'|X_U,\theta)}{q(Y_U')}dY_U' \\
&logp(X_U|\theta) - KL[q(Y_U')||p(Y_U'|X,\theta)] \\
&= l(\theta) - KL[q(Y_U')||p(Y_U'|X,\theta)]
\end{aligned}
\tag{5.9}
$$

The meaning of the last decomposition in the Eq.5.9 is that, for any fixed θ, the free energy term has an upper bound by the log likelihood of the data $l(\theta)$ because KL can never be negative. In order to reach that upper bound when θ is fixed, we need to minimise $KL[q(Y_U')||p(Y_U'|X,\theta)$. The KL distance has its minimum value at zero only if $q(Y_U')$ is equal to $p(Y_U'|X,\theta)$. Therefore, for given fixed model parameters, we can reach the upper limit of the lower bound, by simply replacing the arbitrary function of latent variable as the current estimated posterior of the latent variable:

$$q(Y_U') = p(Y_U'|X_U,\theta) \tag{5.10}$$

The above Eq.5.10 is actually the E-step and pseudo labelling in Eq.5.5 is doing exactly the same thing as in Eq.5.10.

We now explain how M-step increases the log likelihood too. We need to rewrite the log of data likelihood by combing Eq.5.8 and Eq.5.9:

$$l(\theta) = \underbrace{\mathscr{F}(q(Y_U'),\theta)}_{lower\ bound} + KL[q(Y_U')||p(Y_U'|X,\theta)] \tag{5.11}$$

The subsequent M-step is applying supervised learning to optimise the model parameters with fixed pseudo labels which are produced from the precursory E-step. As supervised learning can be seen as maximium likelihood estimation, thereby, M-step increases the likelihood as if the latent variables were observed, resulting in rising the data likelihood's lower bound [46]. In the above Eq.5.11, the $q(Y_U')$ is fixed but estimated from the old parameters θ^{n-1} whereas the posterior of the latent variable is now updated as $p(Y_U'|X,\theta^n)$, therefore they are not equal anymore and the KL term will become a positive value. Together, it is easy to tell that the M-step increases the data log likelihood by at least the increased amount of the lower bound.

Until now, it is clear to see that, how the pseudo labelling (E-step), combined with semi-supervised optimisation of model parameters (M-step) can never decrease the likelihood of the data, leading to guaranteed convergence towards local optima [46]. In summary, the entire process of increasing data log likelihood can be expressed as in Eq.5.12:

$$\underbrace{l(\theta^n) \geq}_{Jensen's\ ineq.} \underbrace{\mathscr{F}(q(Y_U')^n,\theta^n) \geq}_{M-step} \underbrace{\mathscr{F}(q(Y_U')^n,\theta^{n-1}) = l(\theta^{n-1})}_{E-step} \tag{5.12}$$

5.5 Generalisation of Pseudo Labels via Variational Inference for Segmentation

In the last section 5.4, we use an empirical estimation of the posterior of the latent variables (pseudo labels) by setting the T as 0.5. The fixed empirical estimation of T could be sub-optimal especially in the early stage of training when the networks do not have good representations and the predictions are not very confident [140]. Potentially, noisy training with some "bad" pseudo labels could accumulate some errors into the learnt representations. To address this potential issue, we provide an alternative approach to learn to approximate the true posterior of the pseudo labels. This alternative approach can be seen as a generalisation of the empirical estimation approach in SegPL in section 5.4.

5.5.1 Confidence threshold as latent variable

In the last section 5.4, we directly treat pseudo labels as latent variables. However, in the segmentation task, the pseudo labels are pixel-wise, making the generative task a difficult one. To address this, we now introduce a simplification of the graphical model of the pseudo-labelling in 5.4. The key of this simplification is to treat the threshold value T as the latent variable which is a single value for each image:

$$p(X_U|\theta) = \int p(X_U,T|\theta)dT \tag{5.13}$$

This simplification makes the computation of the posterior much easier. Firstly, we have a prior knowledge of the range of this single value T, that any distribution describing values between 0 and 1 can be used as a prior distribution to approximate the real distribution of T. Secondly, the approximation of a single value T is intrinsically simpler than the approximation of pixel-wise unknown labels Y'. To see why the approximation of the true posterior of Y' is very difficult one, we write down the posterior of T with Bayes' rule:

$$p(T|X_U,\theta) = \frac{p(X_U|T,\theta)p(T)}{p(X_U|\theta)} \tag{5.14}$$

The new E-step at iteration n with threshold as the latent variable now becomes:

$$\begin{gathered} p(T_n = i|X_U,\theta^{n-1}) = \\ \frac{\prod_{u=L+1}^{N} p(x_u|\theta^{n-1},T_n = i)p(T_n = i)}{\sum_{j\in[0,1]}\prod_{u=L+1}^{N} p(x_u|\theta^{n-1},T_n = j)p(T_n = j)} \end{gathered} \tag{5.15}$$

From the above Eq.5.15, one can tell that the empirical estimation of the threshold T is actually necessary although not optimal. Even after a major simplification of the latent variables, the posterior of the pseudo-labels is still intractable. Because there are infinite possible values between 0 and 1 in the denominator in Eq. 5.15.

5.5.2 Variational E-step

To address the aforementioned intractable issue in Eq. 5.15, we use variational inference for the approximation of $p(T)$. Luckily, after the above simplification of the graphical model, we actually have a clear prior of T to match, which is any distribution describing values between 0 and 1. For the implementation simplicity, we adapt a univariate Normal distribution for the prior distribution and we denote the prior distribution of T as a surrogate distribution $q(\beta)$. We use extra model parameters ϕ to parameterize the log variance and the mean of the approximated posterior distribution of T, conditioning on the image features, see the beneath Eq. 6.3. ϕ is implemented as a average pooling layer followed by a single 3 x 3 convolutional block including ReLU and normalisation layer, then two 1 x 1 convolutional layers for μ and $Log(\sigma^2)$ respectively. Alternatively, a simple fully connected layer can also be used as ϕ, we found no performance differences among different choices of architectures for ϕ.

$$(\mu, Log(\sigma^2)) = \phi(\theta(X_U)) \tag{5.16}$$

$$p(T|X_U, \theta, \phi) \approx \mathcal{N}(\mu, \sigma) \tag{5.17}$$

Differing from the fixed T in E-step in Eq. 5.4, the T in the variational E-step is dynamic, we denote the stochastic threshold as $\mathbf{T}$ for clarity. We use the standard reparameterization trick [154] to generate the threshold in each iteration:

$$\begin{aligned} &\mathbf{T} = \mu + rand * e^{0.5*log(\sigma^2)} \\ &rand \sim \mathcal{N}(0,1) \end{aligned} \tag{5.18}$$

As demonstrated in previous Eq.5.11 that the data likelihood term has an Evidence Lower Bound (ELBO) which contains a conditional probability of the data given latent variable and a KL distance between the posterior and the prior of the latent variable. We therefore write down variational unsupervised learning objective as:

$$\begin{aligned} &Log(P(X_U) \geq \\ &\sum_{u=L+1}^{N} \mathbb{E}_{T\sim P(\mathbf{T})}[Log(P(x_u|\mathbf{T}))] - KL(p(\mathbf{T})||q(\beta)) \end{aligned} \tag{5.19}$$

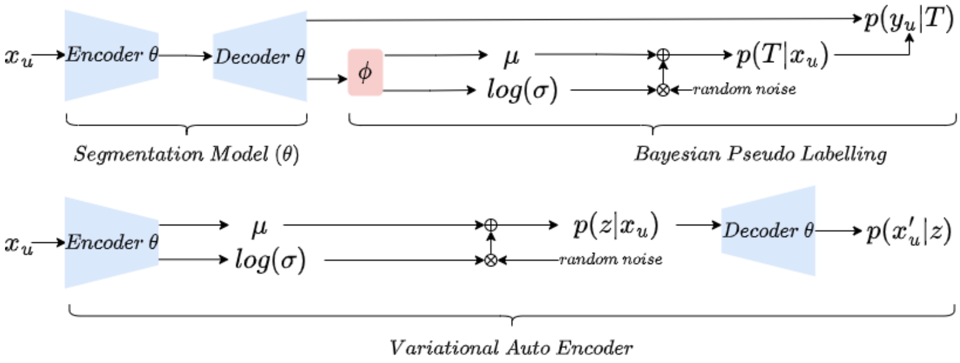

Figure 5.3: Comparison between BPL and VAE in details. For BPL, only the unsupervised learning part is illustrated.

Loss function of BPL The new learning objective $P(X, \mathbf{T}, \theta)$ over the whole data set has a supervised learning $P(X_L, \mathbf{T}, \theta)$ which has not changed from Eq. 5.7, and an unsupervised learning part $P(X_U, \mathbf{T}, \theta)$ from the above Eq. 5.19. The final loss function is an ELBO over the whole data set:

$$\begin{gathered}\mathscr{L}^{VI}_{SegPL} = \underbrace{\frac{1}{L}\sum_{l=1}^{L} f_{dice}(\theta^{n-1}(x_l), y_l) +}_{\mathscr{L}_L} \\ \underbrace{\alpha\frac{1}{N-L}\sum_{u=L+1}^{N} f_{dice}(\theta^{n-1}(x_u), \mathbb{1}(\theta^{n-1}(x_u) > \mathbf{T})) +}_{\mathscr{L}_U} \\ \underbrace{Log(\sigma_\beta) - Log(\sigma) + \frac{\sigma^2 + (\mu - \mu_\beta)^2}{2 * (\sigma_\beta)^2} - 0.5}_{\mathscr{L}_{KL}: KL(p(T)||q(\beta)), \beta \sim \mathscr{N}(\mu_\beta, \sigma_\beta)}\end{gathered} \tag{5.20}$$

Where $\mathbf{T}$ can be found in Eq. 5.18. Different data sets might need different priors for the best empirical performances. For example, we found that $\beta \sim \mathscr{N}(0.4, 0.1)$ works well for CARVE. For BRATS, we didn't tune the prior and simply adopted $\beta \sim \mathscr{N}(0.5, 0.1)$.

5.6 The connection between Bayesian Pseudo Label and Variational Autoencoder

We would like to clarify that the proposed BPL is different from a Variational Auto Encoder (VAE) [155]. First of all, our BPL and VAE implement the latent variables differently as shown in Fig.5.3. Also, VAE has a likelihood term on the data reconstruction $p(x')$, while BPL estimates the likelihood $p(y')$ of the label of unlabelled data. In addition, VAE is fully unsupervised while BPL is semi-supervised.

5.7 Experimental Results

5.7.1 Data sets

The classification of pulmonary arteries and veins (CARVE) We use CARVE for demonstration of 3D binary segmentation of lung vessel of CT images. The CARVE dataset [127] comprises 10 fully annotated non-contrast low-dose thoracic CT scans. Each case has between 399 and 498 images, acquired at various spatial resolutions ranging from (282 x 426) to (302 x 474). We randomly select 1 case for labelled training, 2 cases for unlabelled training, 1 case for validation and the remaining 5 cases for testing. All image and label volumes were cropped to $176 \times 176 \times 3$. To test the influence of the number of labelled training data, we prepared four sets of labelled training volumes with differing numbers of labelled volumes at: 2, 5, 10, 20. Normalisation was performed at case wise. Data curation resulted in 479 volumes for testing, which is equivalent to 1437 images. No data augmentation is used.

BRATS 2018 We use BRATS 2018 [128] for demonstration of 2D multi-class segmentation of brain tumour of MRI images. The BRATS 2018 comprises 210 high-grade glioma and 76 low-grade glioma MRI cases. Each case contains 155 slices. We focus on multi-class segmentation of sub-regions of tumours in high grade gliomas (HGG). All slices were centre-cropped to 176 x 176. We prepared three different sets of 2D slices for labelled training data: 50 slices from one case, 150 slices from one case and 300 slices from two cases. We use another 2 cases for unlabelled training data and 1 case for validation. 50 HGG cases were randomly sampled for testing. Case-wise normalisation was performed and all modalities were concatenated. A total of 3433 images were included for testing. No data augmentation is used.

Task01 Brain Tumour We use Task01 Brain Tumour from Medical Segmentation Decathlon consortium [131] as a demonstration of 3D binary segmentation of brain tumour of MRI images. The Task01 Brain Tumour is based on BRATS 2017 with different naming format from BRATS 2018. This data set was not in our previous MICCAI version but we included this data set here because it is easy to download and use for the readers for the future follow-up works. Each case in The Task01 Brain Tumour has 155 slices with 240 x 240 spatial dimension. We merge all of the tumour classes into one tumour class for simplicity. We do not apply centre cropping in the pre-processing here. In the training, we randomly crop volumes on the fly with size of 64 x 64 x 64. We separate the original training cases as labelled training data and testing data. We use the original testing cases as unlabelled data. For the labelled training data, we use 8 cases with index number from 1 to 8. We have 476 cases for testing and 266 cases for unlabelled training data. We apply normalisation with statistics of intensities across the whole training data set. We keep all of the MRI modalities as 4 channel input.

5.7.2 Baselines

Our baselines include both supervised and semi-supervised learning methods. We use U-net [60] in SegPL as an example of a segmentation network. Partly due to computational constraints, for 3D experiments we used a 3D U-net with 8 channels in the first encoder such that unlabelled data can be included in the same batch. For 2D experiments, we used a 2D U-net with 16 channels in the first encoder. The first baseline utilises supervised training on the backbone and is trained with labelled data denoted as “Sup”. We compared SegPL with state-of-the-art consistency based methods: 1) “cross pseudo supervision” or CPS [139], which is considered the current state-of-the-art for semi-supervised segmentation; 2) another recent state-of-the-art model “cross consistency training” [4], denoted as “CCT”, due to hard-

Table 5.1: Hyper-parameters used across experiments. Different data might need different α.

Data	Batch Size	Learning rate	Steps	α	Unlabelled/labelled
BRATS(2D)	2	0.03	200	0.05	5
CARVE(3D)	2	0.01	800	1.0	4
Task01(3D)	1	0.0004	25000	0.1	2

ware restriction, our implementation shares most of the decoders apart from the last convolutional block; 3) a classic model called "FixMatch" (FM) [2]. To adapt FixMatch for a segmentation task, we added Gaussian noise as weak augmentation and "RandomAug" [133] for strong augmentation; 4) "self-loop [156]", which solves a self-supervised jigsaw problem as pre-training and combines with pseudo-labelling.

5.7.3 Training

We use Adam optimiser[58] with default settings. Our code is implemented using Pytorch 1.0[132] and released in `https://github.com/moucheng2017/EMSSL`. We trained all of the experiments with a TITAN V GPU with 12GB memory. The training hyperparameters are included in Table 5.1.

5.7.4 Pre processing Of the Labels For Multi-Class Segmentation

The pre-processing of the labels of BRATS has two steps: 1) label fusion to mitigate the severe class imbalance between the minority tumour classes and the majority background healthy tissue class; 2) turning a multi-class label into multiple binary labels, for each binary prediction, we can use Sigmoid followed by confidence thresholding for pseudo labelling, combined with separate argmax operation when labels overlap. Here we show two examples of label pre-processing, one is an abstract example (Fig.5.4) and the other one is a real example (Fig.5.5).

5.7.5 Segmentation performances

The segmentation performances of CARVE 2014, BRATS 2018, Task 01 can be found in Tab.5.2, Tab.5.3 and Tab.5.4, respectively. As reflected in the quantitative results in tables, pseudo labelling based SegPL consistently achieves better results than the baselines of semi-supervised and supervised methods. Especially, as shown

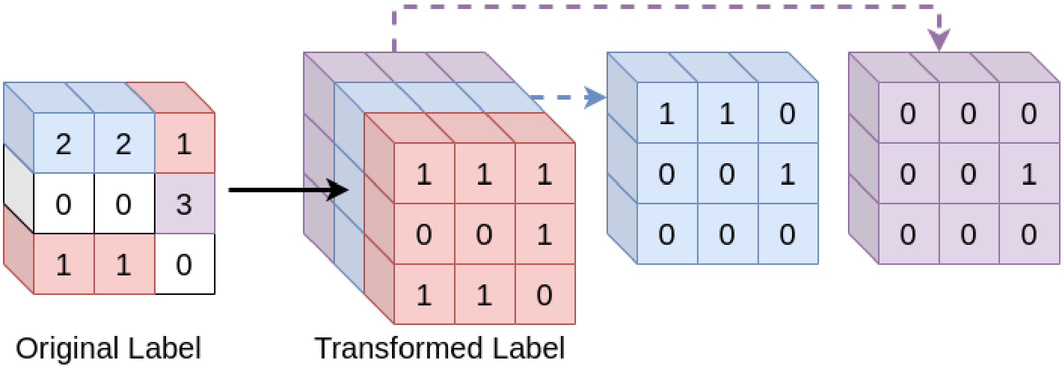

Figure 5.4: An example of the pre-processing of one label of BRATS. 3: enhancing tumour core. 2: tumour core containing enhancing tumour core. 1: whole tumour containing class 2 and 3. 0: healthy tissues. Different colours represent different classes.

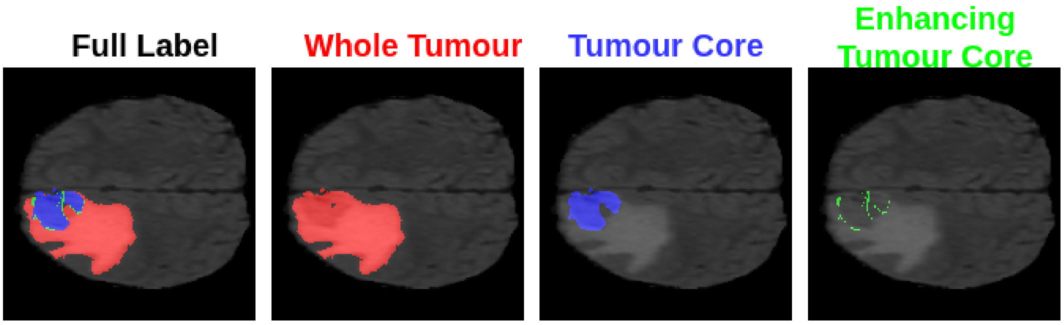

Figure 5.5: Label fusion and binarized labels. Red: whole tumour including tumour core. Blue: tumour core including enhancing tumour core. Green: enhancing tumour core. Segmentation of each tumour class is a binary segmentation.

in Fig.5.6 of the Bland-Altman plot between the best performing baseline CPS and our SegPL on CARVE when only 2 labelled volumes are used for training, SegPL statistically outperforms the best baseline. We further confirm the statistical difference by performing Mann Whitney test on the same results on 2 labelled volumes and we found the p-vale less than 1e-4. By extending the SegPL with variational inference to SegPL-VI, we found further improvements on segmentation on most of the experiments. Interestingly, the improvements brought by SegPL-VI is more obvious on multi-class experiments on BRATS 2018. As the outputs on BRATS are multi-channel but SegPL-VI learns one threshold across all of the channel, we suspect that might bring in strong regularisation effect which results in noticeable improvements. We also noticed that SegPL-VI could fail to learn optimal threshold

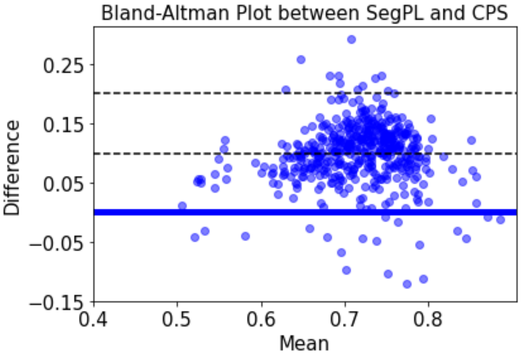

Figure 5.6: SegPL statistically outperforms the best performing baseline CPS when trained on 2 labelled volumes from the CARVE dataset. Each data point represents a single testing image.

sometimes as the result of SegPL-VI on CARVE with 5 labelled volumes are inferior to the corresponding result of SegPL. We expect that more hyper-parameter searching could improve the performance of SegPL.

As shown in the qualitative results in Fig5.7 of CARVE, SegPL successfully learnt better decision boundary than other baselines that SegPL can partially separate the foreground lung vessels from the background whereas most of the other methods classifies everything as background. However, SegPL seemed to have overconfident predictions on the edges of the foreground that it has a lot of false positive results. Similarly in BRATS, SegPL detected one more class of brain tumour (blue) than the other baselines in Fig5.8. However, none of the methods including SegPL can detect the most rare green class of tumour.

One interesting result is shown in Tab.5.4 on 3D binary segmentation of whole tumour. During training, we use random cropping with fixed size at 64 x 64 x 64 to compensate with the memory of GPU. On testing data, we examined the models with different sizes of cropped volumes at 32^3, 64^3, 96^3 and 128^3. The models actually generalise well on the scales that they haven't seen during the training. In fact, larger cropped volumes result in better results.

Although SegPL achieves higher segmentation accuracy, SegPL enjoys a low computational burden. As illustrated in the computational need section in Tab.5.2, SegPL has the least computational burden among all of the tested semi-supervised

Table 5.2: Our model vs Baselines on a binary vessel segmentation task on 3D CT images of the CARVE dataset. Metric is Intersection over Union (IoU (↑) in %). Avg performance of 5 training. blue: 2nd best. red: best

Data	**Supervised**	**Semi-Supervised**				
Labelled Volumes	**3D U-net** [60](2015)	**FixMatch** [2](2020)	**CCT** [4](2020)	**CPS** [139](2021)	**SegPL** (Ours, 2022)	**SegPL+VI** (Ours, 2022)
2	56.79±6.44	62.35±7.87	51.71±7.31	66.67±8.16	**69.44±6.38**	**70.65±6.33**
5	58.28±8.85	60.80±5.74	55.32±9.05	70.61±7.09	**76.52±9.20**	**73.33±8.61**
10	67.93±6.19	72.10±8.45	66.94±12.22	75.19±7.72	**79.51±8.14**	**79.73±7.24**
20	81.40±7.45	80.68±7.36	80.58±7.31	81.65±7.51	**83.08±7.57**	**83.41±7.14**
Computational need						
Train(s)	1014	2674	4129	2730	1601	1715
Flops	6.22	12.44	8.3	12.44	6.22	6.23
Para(K)	626.74	626.74	646.74	1253.48	626.74	630.0

Table 5.3: Our model vs Baselines on multi-class tumour segmentation on 2D MRI images of BRATS 2018. Metric is Intersection over Union (IoU (↑) in %). Avg performance of 5 runs. blue: 2nd best. red: best

Data	**Supervised**	**Semi-Supervised**				
Labelled Slices	**2D U-net** [60](2015)	**Self-Loop** [156](2020)	**FixMatch** [2](2020)	**CPS** [139](2021)	**SegPL** (Ours, 2022)	**SegPL+VI** (Ours, 2022)
50	54.08±10.65	65.91±10.17	67.35±9.68	63.89±11.54	**70.60±12.57**	**71.20±12.77**
150	64.24±8.31	68.45±11.82	69.54±12.89	69.69±6.22	**71.35±9.38**	**72.93±12.97**
300	67.49±11.40	70.80±11.97	70.84±9.37	71.24±10.80	**72.60±10.78**	**75.12±13.31**

Table 5.4: Our model vs Supervised baseline on 3D binary tumour segmentation of Task 01 Brain Tumour (BRATS 2017). Metric is Intersection over Union (IoU (↑) in %). Avg performance of models between iteration 20000 and 25000 with 1000 as the interval. red: best

Testing size	32^3	64^3	96^3	128^3
Supervised	61.07±7.93	66.94±12.4	70.13±13.22	72.09±12.48
SegPL-VI	64.44±8.3	71.43±11.91	73.07±11.71	74.48±11.51

learning baselines. Especially in terms of FLOPs, SegPL is very close to supervised learning methods. This shows that our model has the scaling potential for large models and large data sets.

5.7.6 Ablation studies on hyper-parameters

We performed brief ablation studies on hyper-parameters on BRATS with 150 labelled slices. As shown in Fig.5.9, a) shows that SegPL is very sensitive to learning rate that it should be at least 0.01. We found that other baselines also needed a large learning rate. Fig.5.9.b) shows the impact of a warm-up schedule of α from 0 to final α value. x axis is the length of linear warming-up of α in terms of whole steps.

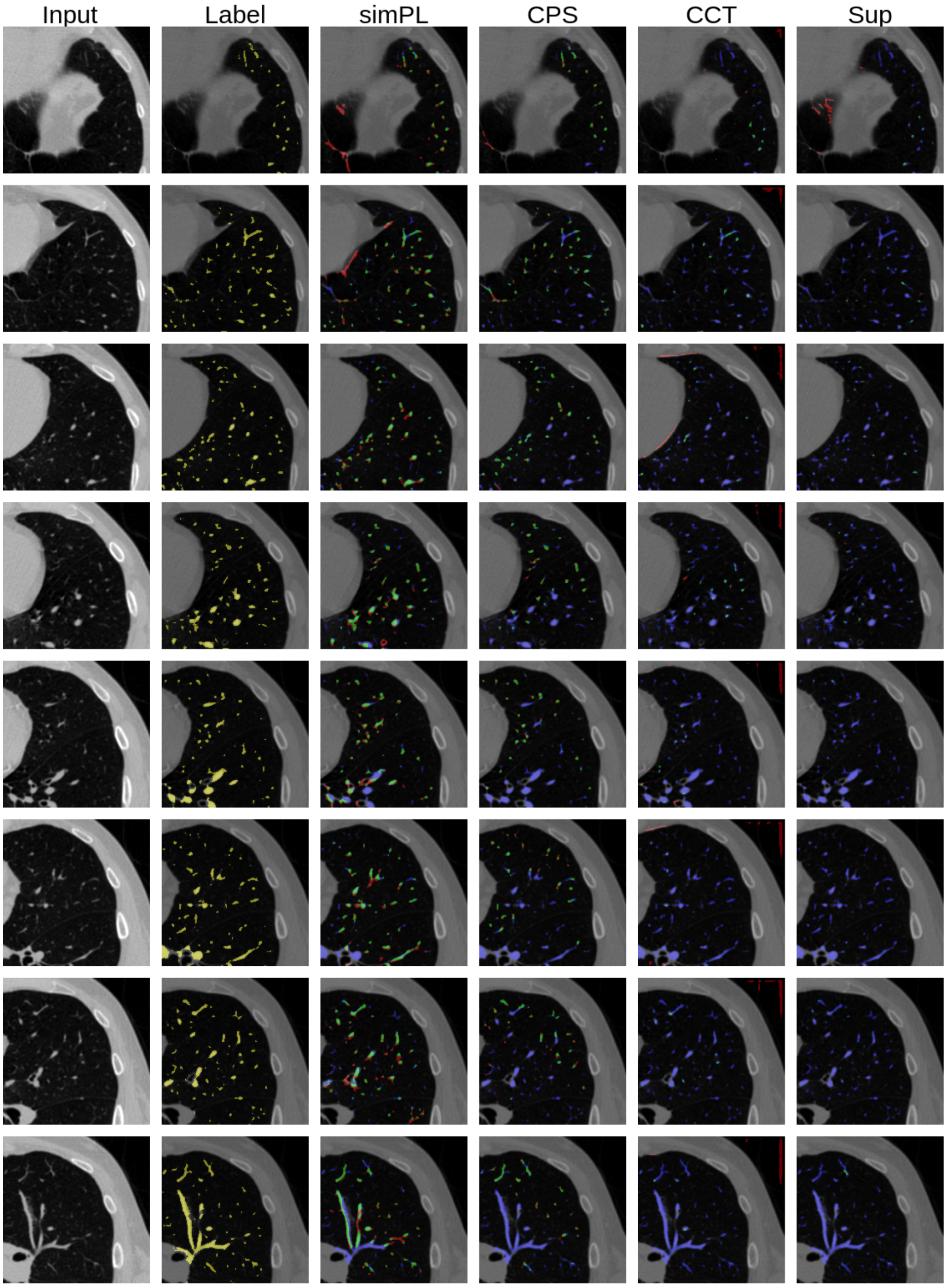

Figure 5.7: Visual results. CARVE trained with 5 labelled volumes. Red: false positive. Green: true positive. Blue: false negative. Yellow: ground truth. GT: Ground truth. CPS: cross pseudo labels (CVPR 2021). CCT: cross consistency training (CVPR 2020). Sup: supervised training.

It appears that SegPL is not sensitive to the warm-up schedule of α. Fig.5.9.c) illustrates the effect of the ratio between unlabelled images to labelled images in each batch. The suitable range of unlabelled/labelled ratio is quite wide and between 1 to 10. Fig.5.9.d) shows that the pseudo supervision cannot be too strong.

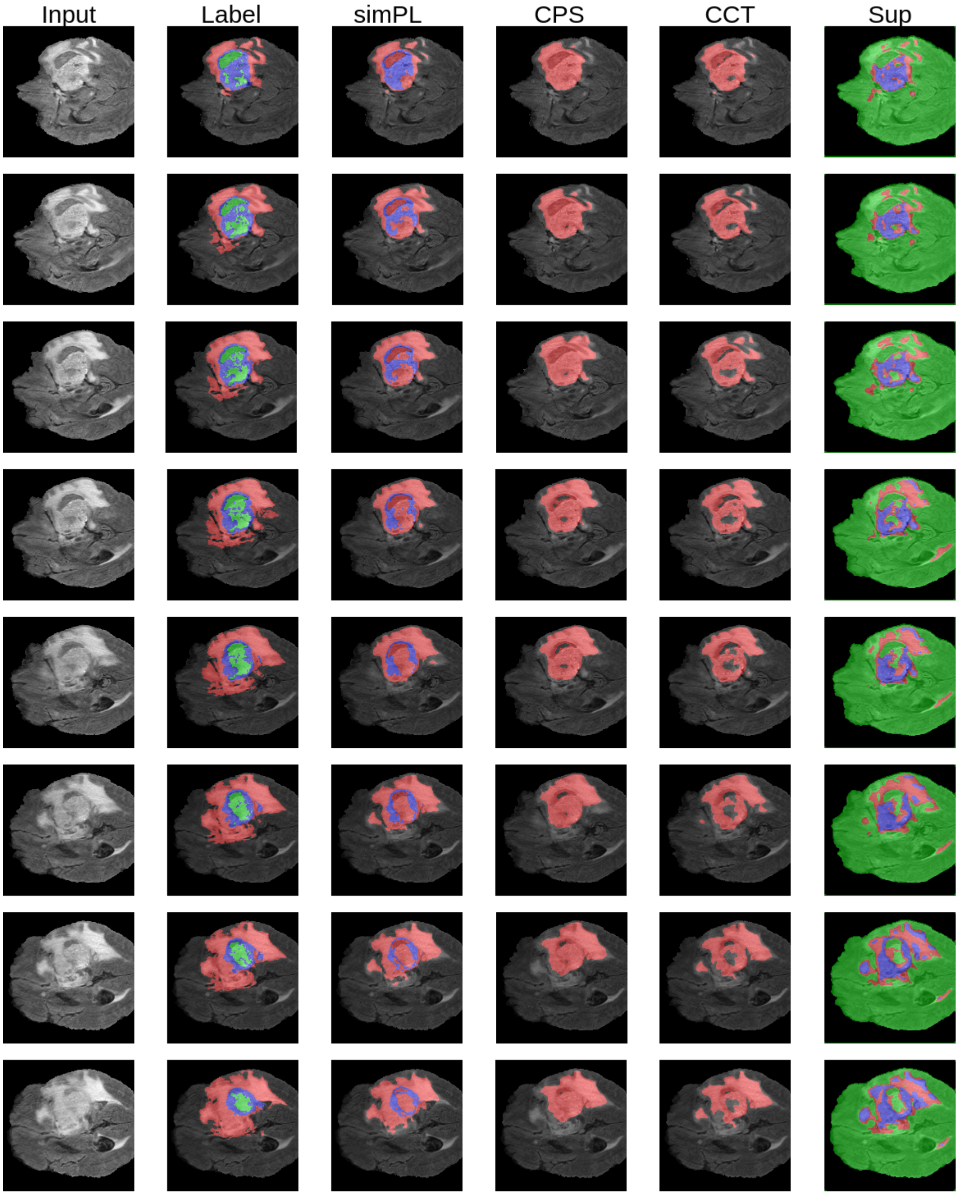

Figure 5.8: Visual results. BRATS 2018 trained with 300 labelled slices. Red: whole tumour. Green: tumour core. Blue: enhancing tumour core. GT: Ground truth. CPS: cross pseudo labels (CVPR 2021). CCT: cross consistency training (CVPR 2020). Sup: supervised training.

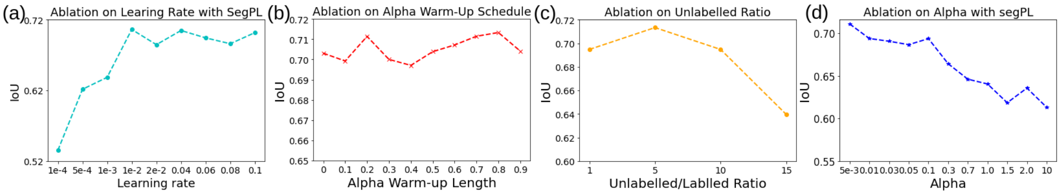

Figure 5.9: Ablation studies on BRATS with 150 labelled slices.

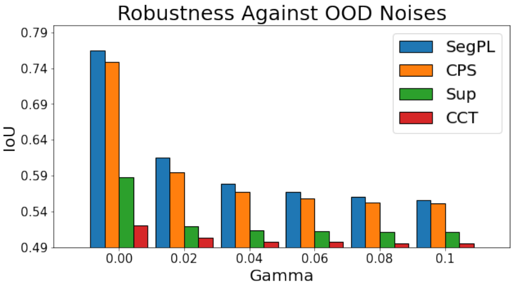

Figure 5.10: Robustness against out-of-distribution noise. Gamma is the strength of the out-of-distribution noises. Using 2 labelled volumes from CARVE.

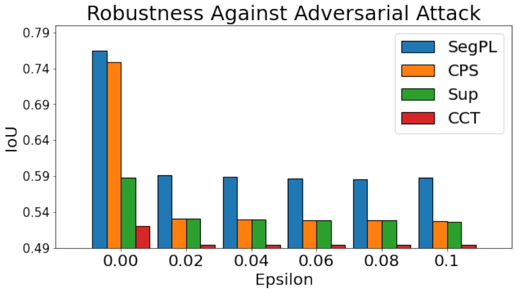

Figure 5.11: Robustness against adversarial attack. Epsilon is the strength of the FGSM[6] attack. Using 2 labelled volumes from CARVE.

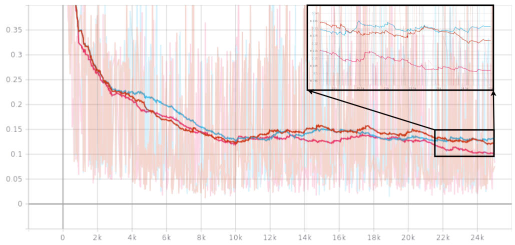

Figure 5.12: Ablation studies on priors of 0.4, 0.5 and 0.7. Y-axis: segmentation losses. X-axis: training iterations. Red: prior = 0.5. Blue: prior = 0.4. Pink: prior = 0.7. All trained on Task 01 Brain Tumour.

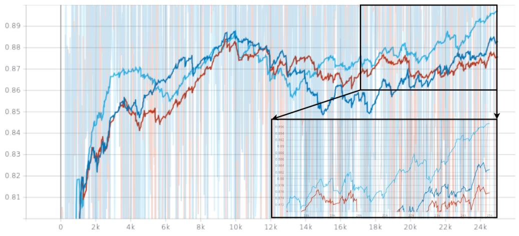

Figure 5.13: Ablation studies on priors of 0.4, 0.5 and 0.7. Y-axis: segmentation IoUs. X-axis: training iterations. Red: prior = 0.5. Blue: prior = 0.7. Dark blue: prior = 0.5. All trained on Task 01 Brain Tumour.

This confirms the suggestions from the original pseudo labelling paper that pseudo supervision should not dominate the training.

5.7.7 Ablation studies on the prior of Bayesian Pseudo Labels

According to the result of the ablation study on the prior values in Fig.4.12 and Fig.4.13, we might suggest to the future users to start from a high value starting from 0.9.

5.7.8 Robustness

Medical imaging normally suffer from out-of-the-distribution noises (e.g. variations in scan acquisition parameters and different patient populations) which significantly degenerate the trained model in real-life deployments. We investigate the robustness of SegPL on out-of-distribution (OOD) noise using models trained on CARVE. OOD noises are simulated with unseen random contrast and Gaussian noise, we then apply mix-up [157] to create new testing samples by adding the OOD noises on original images. Specifically, for a given original testing image x_t, we applied random contrast and noise augmentation on x_t to derive OOD samples x_t'. We arrived at the testing sample ($\hat{x_t}$) via $\gamma x_t' + (1-\gamma)x_t$. As shown in Fig.5.10, as testing difficulty increases, the performances across all baselines drop exponentially. SegPL outperformed all of the baselines across all of the tested experimental settings. The findings suggest that SegPL is more robust when testing on OOD samples and achieves better generalisation performance against that from the baselines.

We also examined SegPL's robustness against adversarial attack as privacy-preserving collaborative federated learning among hospitals has now become popular. We focus on using a fast gradient sign method (FGSM)[6]. The FGSM works by perturbing the input data by adding a small amount of noise, in the direction of the gradient of the cost function with respect to the input. The objective is to maximize the loss function, causing the model to make incorrect predictions. With increasing strength of adversarial attack (Epsilon), all the networks suffered performance drop. As shown in Fig.5.11, SegPL yet suffered much less than the baselines.

5.7.9 Uncertainty

Since SegPL-VI is trained with a stochastic threshold for unlabelled data therefore not suffering from posterior collapse. Consequently, SegPL can generate plausible segmentation during inference using stochastic thresholds. To test the performance of SegPL-VI on uncertainty quantification, we use random latent variable values (threshold) with 5 Monte Carlo samples. We focus experimenting on models trained with 5 labelled volumes of CARVE data set. For comparison, we adopt Deep Ensemble, as it is the gold-standard baseline for uncertainty estimation [158][159]. Both the tested methods Deep Ensemble and SegPL-VI achieved the same Brier score at 0.97. This result shows that SegPL-VI has the potential to become a benchmark method for uncertainty quantification. The Brier score is calculated using beneath equation Eq.5.21, where, y_{ij} is the ground truth label at pixel at location i,

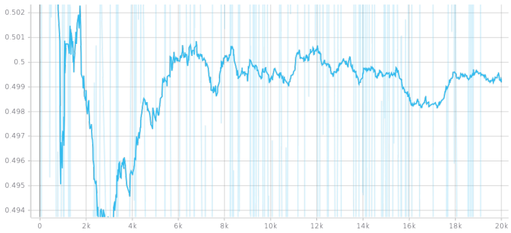

Figure 5.14: Learnt threshold with prior at 0.5 trained on Task 01 Brain Tumour.

j, y_{ij} is 1 for foreground pixel and y_{ij} is 0 for background pixel. p_{ij} is the predicted probability of the pixel being the foreground pixel.

$$Brier = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(p_{ij} - y_{ij})^2 \tag{5.21}$$

5.8 Conclusions

In this paper, we interpret pseudo-labelling as EM and explore the potential improvement with variational inference following generalised EM. We hope this interperation can shed new lights on explainable AI. Empirically, we examined that the original pseudo-labelling [67] on semi-supervised medical image segmentation and we report that pseudo-labelling as a competitive and robust baseline. In the future pipeline for learning with limited annotations, we expect to exploits full potential if we combine semi-supervised learning with large-scale pre-training techniques.

Chapter 6

Deep Variational Parameter Mapping: Applications of Unsupervised Representation Learning to MRI parameter estimation

This chapter presents a deep variational clustering approach for learning without any labels for estimating diffusion MRI parameters. This chapter is based on a manuscript which is in preparation for a submission towards a technical conference. I conceived the methodology, implemented the code, performed the experiments and wrote the draft of the manuscript. My colleague Paddy Slator provided the original code of the baseline and the data simulation. My colleague Toby provided feedback on the mathematical formulations. All of the co-authors contributed to the writing of the manuscript.

6.1 Abstract

We introduce and demonstrate a new paradigm for quantitative parameter mapping in MRI. Parameter mapping techniques, such as diffusion MRI (dMRI) and quantitative MRI (qMRI), have the potential to robustly and repeatably measure biologically-relevant tissue maps that strongly relate to underlying microstructure. Quantitative maps are calculated by fitting a model to multiple images, e.g. with least-squares or machine learning. However, the overwhelming majority of model fitting techniques assume that each voxel is independent, ignoring any co-dependencies in the data. This makes model fitting sensitive to voxelwise measurement noise, hampering reliability and repeatability. We propose a self-supervised deep variational autoencoder model fitting approach that breaks the assumption of independent pixels, leveraging redundancies in the data to effectively perform data-driven regularisation of quantitative maps. We demonstrate that our approach outperforms current model fitting techniques in dMRI simulations and real data. Our approach enables improved quantitative maps and/or reduced acquisition times, and can hence support the clinical adoption of parameter mapping methods such as dMRI and qMRI.

6.2 Introduction

Multiple MRI techniques can produce quantitative maps of biophysical, chemical and physiological tissue properties. Such *quantitative parameter mapping* techniques include diffusion MRI (dMRI) and quantitative MRI (qMRI). dMRI and qMRI use an essentially identical approach; by acquiring multiple images then fitting a model to the images, intrinsic values of the relevant tissue properties in each voxel can be estimated. In dMRI the images have different diffusion weightings and directions, and model fitting enables estimation of microstructural parameters, such as diffusivity and kurtosis. In qMRI acquisition parameters such as echo time or inversion time are varied, and model fits produce maps of chemical tissue properties, such as T1 and T2.

In the vast majority of applications, such models are fit to the data using non-linear least squares techniques. Machine learning model fitting is emerging as a attractive alternative technique. Supervised learning has been demonstrated for a range of models [160], but the distribution of parameters in the training dataset introduces biases [161, 162]. Self-supervised learning has the potential to address this, but has only been demonstrated in a few models thus far, most prominently the intravoxel incoherent motion (IVIM) model [163]. Hybrid approaches are also emerging that merge the benefits of supervised and self-supervised learning [162].

However, whilst these approaches offer improvements, the current generation of parameter mapping techniques fail to capitalize on the extensive inherent redundancies in the data. Specifically, the overwhelming majority of techniques fit models to each voxel separately, effectively assuming that each voxel is independent. This leads to high sensitivity to voxelwise measurement noise, which negatively affects the quality of derived parameter maps. Bayesian hierarchical modelling has been proposed as an approach that breaks these assumptions, but requires slow Markov chain Monte Carlo inference [164]. Convolutional neural networks (CNNs) have been demonstrated for IVIM fitting [165], but only learn spatial redundancies. One technique assumes a set of underlying tissue components to regularise quantitative maps [166] in a data-driven way, but at the expense of voxelwise parameter estimates.

In this paper, we demonstrate a deep learning approach that breaks the paradigm of independent voxels. Analogously to recent approaches [166], we seek a lower dimensional representation of the data to parameterise date redundancies. We therefore adapt ideas from the clustering literature [167] to derive a deep variational autoencoder for quantitative parameter mapping. We show that our approach yields improved parameter estimates and maps in simulated and real data.

6.3 Methods

Whilst we emphasise that our approach is a general solution for quantitative parameter mapping, for simplicity we assume a dMRI dataset and model throughout the Methods. Figure 6.1 is a schematic of our method.

6.3.1 Problem formulation

We assume an observed *discrete* series of, T, dMRI images, $\mathscr{S} = S_{(1)}, S_{(2)}, ..., S_{(T)}$, where $S_{(i)}$ is a diffusion weighted image (DWI) with height, H, width, W, and depth, D. The DWIs are defined $S_{(i)} = \{s_{(i)}^{(h,w,d)}; h \in [1,2,..,H], w \in [1,2,..,W], d \in$

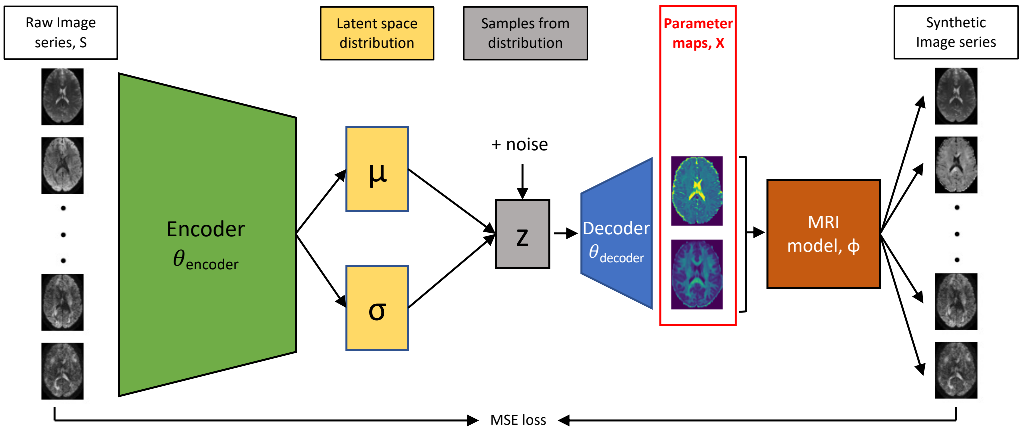

Figure 6.1: A schematic of one of our methods with univariate Gaussian Prior. The quantitative mappings of interest are indicated by the red box. The total loss is MSE loss plus KL loss.

$[1,2,..,D]\}$, where $s_{(i)}^{(h,w,d)}$ is the signal at the voxel (h,w,d) for the ith DWI. We aim to estimate the tissue properties, i.e. model parameters, $X = \{\mathbf{x}^{(h,w,d)}; h \in [1,2,..,H], w \in [1,2,..,W], d \in [1,2,..,D]\}$ where the tissue properties for a voxel located at (h,w,d) corresponds to the vector $\mathbf{x}^{(h,w,d)} = (x_1^{(h,w,d)},...,x_M^{(h,w,d)})$ for the M tissue properties. The model parameters are estimated with respect to the signal model. Traditionally, each voxel would be estimated independently of all the other voxels, for example, the MLE at (h_1,w_1,d_1) is calculated by finding the parameters that maximise the probability of seeing the observed data; i.e. what maximises $p((s_{(1)}^{(h_1,w_1,d_1)},...,s_{(T)}^{(h_1,w_1,d_1)})|\mathbf{x}^{(h_1,w_1,d_1)})$. This is performed independently to the MLE at (h_2,w_2,d_2). This obviously does not consider the global information of voxels, and is very inefficient.

6.3.2 Probabilistic model on jointly estimating across all voxels

In this work, we propose to jointly model the distribution of the voxels together as $p_\theta(\mathscr{S})$, which is a computationally challenging task. To address this issue, we propose to project the implicit high-dimensional distribution of $\mathscr{S}$ from the data space into an explicit low-dimensional continuous distribution in a latent space. We therefore need to introduce latent variables, $\mathbf{z}$, we anticipate that these variables will capture underlying biologically-relevant structures in the data. Formally, our goal here is to maximise the following joint distribution: $p_\theta(\mathscr{S}) = \int p_\theta(\mathscr{S}|\mathbf{z})p_\theta(z)dz$. Considering dMRI-specific format, with independence assumption for expression clarity (we note that despite this assumption our proposed model does not treat voxels as independent due to the shared latent space):

$$p_\theta(\mathscr{S}) \approx \prod_{t=1}^{t=T}\prod_{d=1}^{d=D}\prod_{w=1}^{w=W}\prod_{h=1}^{h=H} p_\theta(s_{(t)}^{(h,w,d)}|\mathbf{z})p_\theta(\mathbf{z}) \tag{6.1}$$

Our latent modelling choice in Eq. 6.1 enjoys two benefits. The first benefit is, by conditioning the data from each DWI from each voxel $S_{(i)}^{(h,w,d)}$ on $\mathbf{z}$, we absorb all the arbitrary dependencies among voxels into $\mathbf{z}$, a compact representation in a latent space. In the latent space which has lower dimension than the data space, voxels are clustered with their close voxels, therefore inter-voxels information must be captured. The second benefit is, the complicated underlying distribution of $p_\theta(\mathscr{S})$ can be learnt via learning a much simpler distribution $p_\theta(\mathbf{z})$. However, the marginal distribution of $p_\theta(\mathbf{z})$ is still intractable: $p_\theta(\mathbf{z}|\mathscr{S}) = p_\theta(\mathscr{S},\mathbf{z})/p_\theta(\mathscr{S})$ because the data density is unknown. To address this computational issue, we deploy a variational approach to approximate the posterior of $p_\theta(\mathbf{z})$.

6.3.2.1 Univariate Gaussian prior

We explore different implementations to obtain the posterior of $p(\mathbf{z})$. Let's denote the prior of $\mathbf{z}$ as $q(\mathbf{z})$. We start with a simple univariate Gaussian Prior as $\mathscr{N}(0,1)$. We first use the encoder to parameterize $\mathbf{z}$ from input observed signals:

$$\mu, Log(\sigma^2) = \theta_{Encoder}(\mathscr{S}) \tag{6.2}$$

$$p(\mathbf{z}|\mathscr{S}, \theta_{Encoder}) \approx \mathscr{N}(\mu, \sigma) \tag{6.3}$$

We then use a decoder to map randomly drawn samples $\mathbf{z} \sim p(\mathbf{z})$ to physically-relevant dMRI model parameters (X) that are inputs to a closed-form dMRI model that reconstructs the MRI signal. We denote the closed-form physics decoding process as $\phi(.)$. We emphasise that $\phi(.)$ can be any dMRI (or qMRI) model; the details of the dMRI models we use are described in section 6.3.3. The complete decoding process is:

$$p(\mathscr{S}) = p(\phi(X|\mathbf{z}, \theta_{Decoder})) \tag{6.4}$$

The loss function becomes an evidence lower bound as:

$$Log(p_\theta(\mathscr{S})) \geq \sum_{d=1}^{d=D}\sum_{w=1}^{w=W}\sum_{h=1}^{h=H} \mathbb{E}_{z\sim p(\mathbf{z})}[Log(p(\phi(\mathbf{x}^{(h,w,d)}|\mathbf{z}, \theta)))] - KL(p(\mathbf{z})||q(\mathbf{z})) \tag{6.5}$$

The likelihood of $Logp(\phi(\mathbf{x}^{(h,w,d)}|\mathbf{z}, \theta)))$ is measured as a mean squared error loss between estimated signals and raw input signals.

6.3.2.2 Gaussian Mixture Prior

We further enhance our model's expressivity by considering a prior as a mixture of univariate Gaussians. We add an extra latent variable $\mathbf{y}$ for controlling the index of the Gaussian component. The prior of $\mathbf{y}$ is chosen as a Categorical distribution. The probabilistic model is:

$$p_\theta(\mathscr{S}) = \int_{\mathbf{z}}\int_{\mathbf{y}} p_\theta(\mathscr{S}|\mathbf{z})p(\mathbf{z}|\mathbf{y})p(\mathbf{y})d\mathbf{y}d\mathbf{z} \tag{6.6}$$

In implementation, we build our Gaussian mixture VAE (VAE-GMM) following a hierarchical order. We first need to parametrize the mixing coeffients of each Gaussian using Gumbel Softmax [168, 169]:

$$c = Gumbel(\theta_{Encoder_{1st}}(\mathscr{S})) \tag{6.7}$$

Where c is a normalised vector indicating the weight for each Gaussian and the sum of the c is 1. We then concatenate c with input to parametrize the parameters of Gaussians:

$$\mu_k, Log(\sigma_k^2) = \theta_{Encoder_{2nd}}(\mathscr{S}, c) \tag{6.8}$$

$$p(\mathbf{z}|\mathscr{S}, \theta_{Encoder}) \approx \mathscr{N}(\mu_k, \sigma_k) \tag{6.9}$$

Where k means that the mean and the variance are for the K-th Gaussian. We apply the same decoding process as in the last section. And the loss function now becomes:

$$Log(p_\theta(\mathscr{S})) \geq \mathbb{E}_{z\sim p(\mathbf{y}|\mathbf{z})p(\mathbf{z}|x)}[Log(p(\phi(\mathbf{X}|\mathbf{z},\theta)))] - KL(p(\mathbf{z},\mathbf{y})||q(\mathbf{z},\mathbf{y})) \quad (6.10)$$

6.3.3 MRI models

We test our approach on two dMRI models, the mean signal diffusion kurtosis imaging (MS-DKI)[170] model and ball-stick model [171]. The normalised signal for MS-DKI is given by

$$\phi(b) = \exp\left(-bD + b^2D^2K/6\right) \quad (6.11)$$

where b is the b-value, D is the diffusivity and K the kurtosis. For ball-stick the normalised signal is

$$\phi(b,\mathbf{g}) = f\exp\left(-bD_{||}(\mathbf{g}.\mathbf{n})\right) + (1-f)\exp\left(-bD_{iso}\right) \quad (6.12)$$

where b is the b-value, $\mathbf{g}$ the gradient direction, f is the stick volume fraction, $D_{||}$ is the parallel diffusivity of the stick, and D_{iso} is the ball isotropic diffusivity.

6.3.4 Implementations

We follow [155] and use an auto-encoder architecture for our implementation. Our encoder is 3 fully connected layers and our decoder is one fully connected layer.

6.4 Experimental Results

6.4.1 Baselines

We use least squares fitting and voxel-wise self-supervised fitting as baselines for the real data, and self-supervised fitting as a baseline for the simulated data. We note that voxelwise self-supervised fitting has not been previously demonstrated for the specific MRI models we use. For MSDKI we implemented LSQ fitting with the dipy python package [172], and for ball-stick with the dmipy package [173]. We implemented self-supervised fitting as described in [174]. In summary, we use a three-layer artificial neural network (ANN) with the number of nodes in the first three layers equal to the number of dMRI volumes, and the output layer having the same number of layers as the relevant dMRI model.

6.4.2 Data

We first demonstrate our method on an illustrative toy simulated example with known ground truth. To test our method's ability to capture underlying data redundancies, we simulated an MSDKI dataset with three underlying clusters. We chose each cluster's diffusivity and kurtosis to mimic white matter, grey matter, and CSF; the mean D, K values for each cluster were $\{1, 1.5\}$, $\{1.5, 1\}$, and $\{3, 0\}$ respectively, with diffusivity in units of μm^2/ms. We simulated 10,000 voxels, with relative weightings of each cluster $\{0.5, 0.4, 1\}$. The specific ground truth parameter value was simulated from a Gaussian with the relevant mean D and K for that cluster, and variance 0.1 for white matter and grey matter clusters, and 0.01 for the CSF cluster.

For real data, We use publicly-available dMRI data provided by the HCP WU-Minn Consortium[175] to demonstrate our methods. We used preprocessed[176] data from a single subject from the 1200 Subjects Data Release.

6.5 Results

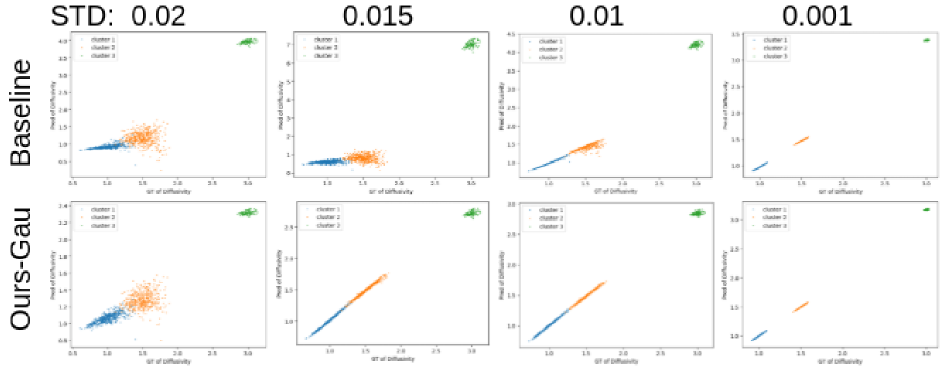

Figure 6.2: Comparisons between self-supervised voxel-wise baseline and ours (Gaussian prior) on simulated model using MSDKI. X axis: ground truth of simulated kurtosis. Y axis: prediction of kurtosis. Ours vastly outperforms the baseline in recovery of the ground truth.

Figure 6.2 demonstrates that our VAE approach significantly outperforms voxelwise self-supervised fitting in the simulated data, as our VAE approach produces predictions aligned better with the ground truth (closer to the diagonal line), with less variances.

We further conducted ablation studies on our VAE model. The first ablation study was to find the most optimal dimension of the latent space and we found that the VAE model couldn't reconstruct the data if the dimension of the latent space is too low as shown in 1st row in Fig.5.9. We also found that the strength (α) of the kl loss has a trade-off as shown in 2nd row in Fig.5.9. However, in general, larger kl loss brings more stochasticity into the modelling, resulting in higher variances shown as more scattered data points.

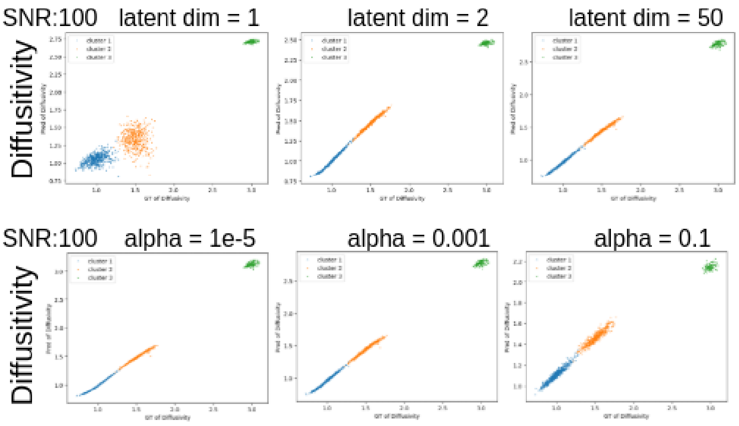

Figure 6.3: Ablation studies. We observe that both latent dimension and kl loss strength have optimal values.

On the real-data set HCP, figure 6.4 shows that VAE improves ball-stick parameter maps compared to the baselines, particularly the stick diffusivity map, which is less noisy and better highlights anatomical structures. We found that Gaussian mixture VAE futher improved the image quality in figure 6.4 with more detailed anatomical structures, for instance, Gaussian mixture VAE successfully capture more details of the white matter comparing all of the other methods (see the stick diffusivity maps and ball diffusivity maps).

However, when the MRI model is much simpler with less parameters such as MSDKI, figure 6.5 shows that our VAE produces maps comparable to the baselines in MSDKI, without producing any hallucinations.

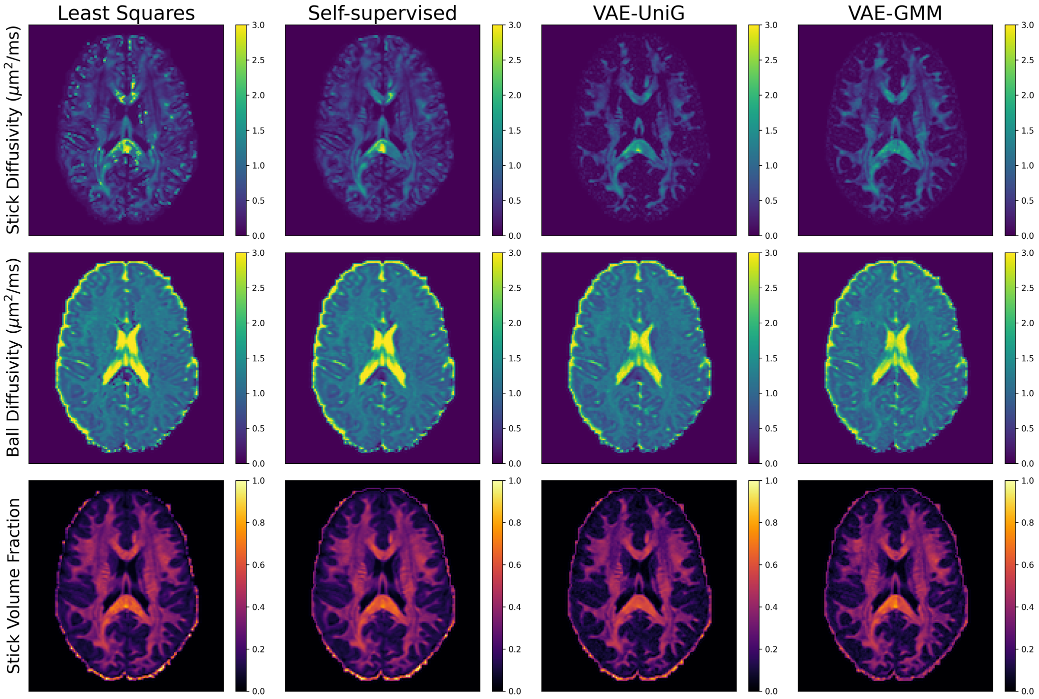

Figure 6.4: LSQ, self-supervised, and VAE ball-stick fits to HCP dMRI subject. Our methods drastically reduces noise and better highlights anatomical features.

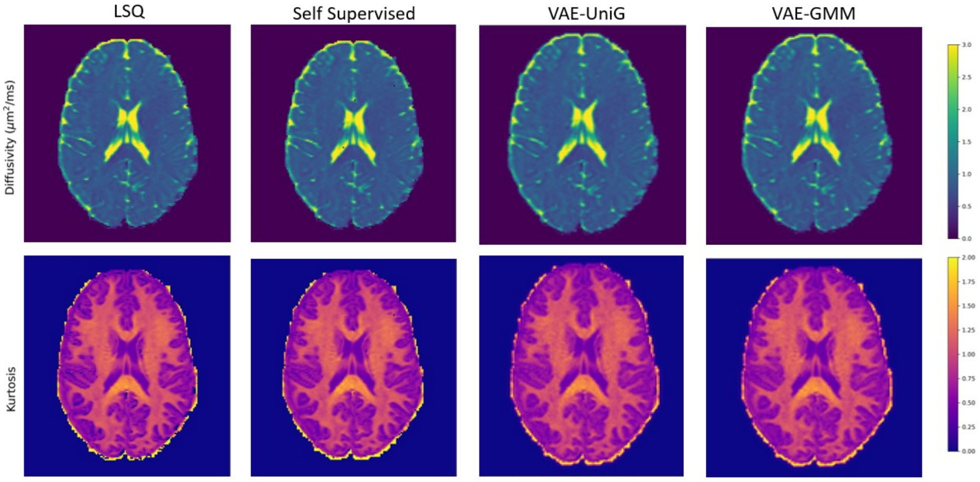

Figure 6.5: Comparisons on MS-DKI fits on HCP dMRI subject. Our approach has less obvious improvements when the MRI model is relatively simple, but doesn't hallucinate spurious anatomical features.

6.6 Discussion on learnt distributions of latent variables

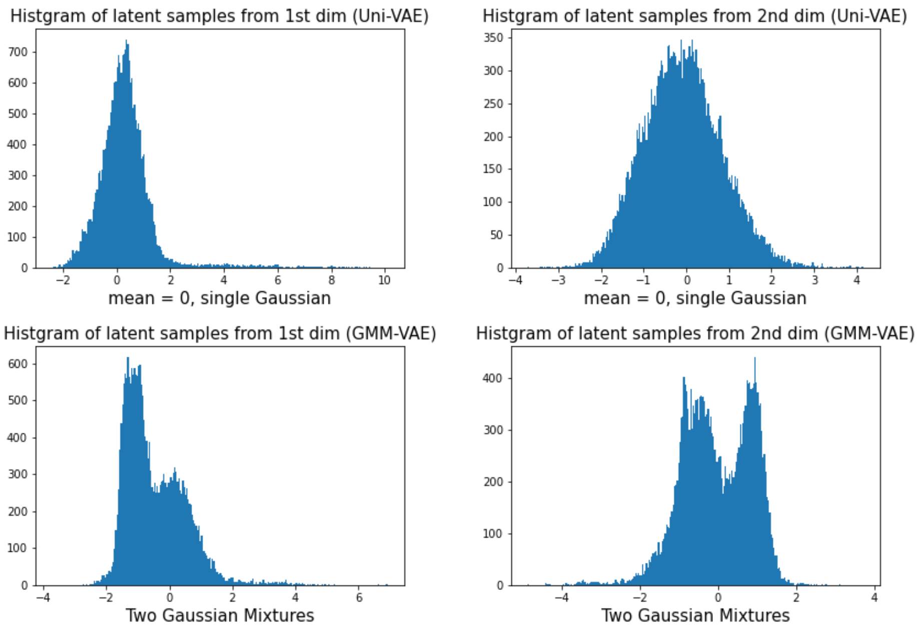

Figure 6.6: Visualisation of the posterior distribution of the latent variable in our VAE models after training on HCP data.

We also studied the learnt posterior distributions and present the results in Fig.6.6. We run stochastic inferences for 500 times for each trained model and we collect the latent z samples and plot their histograms with bin number 250. As shown in Fig.6.6.**Row 1**: VAE with univariate Gaussian prior. In Uni-VAE, we observe that the learnt posterior distribution of the latent variable is similar to Gaussian and it has the mean value at 0 in both dimension 1 and dimension 2, given that the mean value in our priors is also 0. Fig.6.6.**Row 2**: VAE with Gaussian mixture prior. We observe that the learnt posterior distribution of the latent variable is similar to Gaussian Mixture distribution with two obvious peaks. For example in the first dimension of the latent space (row 2 left), the Gaussian mixture has two peaks at around -1 and 0 respectively. In the second dimension of the latent space (row 2 right), the Gaussian mixture has two peaks at around -0.5 and 1. However, we noticed that there are some long tails, implying that there might be more Gaussians, for example see the long tail on the right hand side in 1st dim of GMM-VAE (left bottom of Fig.6.6).

Interestingly, we also noticed that at higher dimensions, the latent space has more complicated patterns which are hard for interpreting, suggesting that the true posterior might not be the Gaussian or mixture of Gaussians.

6.7 Conclusion

We introduce a deep VAE model fitting method that exploits data redundancies to maximise the amount of information extracted in parameter mapping techniques such as dMRI and qMRI. Our deep VAE approach clearly outperforms the baseline methods in simulated and real data with the ball-stick model. Using simulated data with known ground truth, we show that by capturing underlying data structures our VAE approach can significantly improve estimation of ground truth parameter values compared to voxelwise fitting (Figure 6.2). In real data with the ball-stick model, our VAE algorithm reduces noisy features in maps and reveals new anatomical details (Figure 6.4).

We stress that the advantages of our approach can be leveraged in two ways. Firstly, it could be applied to existing dMRI and qMRI acquisition sequences to yield improved quantitative maps of tissue structure and function compared to current fitting techniques, and hence potentially improve the clinical utility of such maps. Alternatively, it could be used to significantly shorten dMRI and qMRI acquisitions, whilst maintaining the quality of derived parameter maps, enabling more comfortable and cheaper scans. Our VAE algorithm can fit any dMRI or qMRI model by simply modifying the closed-form decoder $\phi(.)$.

To conclude, we introduce and demonstrate a deep VAE algorithm for quantitative parameter mapping. Our approach breaks the typical assumption of voxelwise independence, and can hence identify and exploit data redundancies to improve the quality of inferred parameter maps. Our work can enable a new generation of dMRI and qMRI that is more sensitive and specific to underlying tissue structure, and hence support clinical adoption of these techniques.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

7.1.1 Chapter 4

The results from Chapter 4 led to two peer-reviewed publications: one as an oral presentation at MIDL and the other as a journal paper in IEEE TMI. Furthermore, the framework proposed in Chapter 4 received validation for its commercial potential, resulting in a patent application with the assistance of our colleagues at UCLB.

Firstly, Chapter 4 provides a new interpretation of the spatial attention mechanism as an implementation of differential morphological operations on the features. Chapter 4 then explores the application of this new insight in the context of semi-supervised segmentation of medical images, resulting in a framework called MisMatch. MisMatch also addresses the challenge of integrating consistency regularization from semi-supervised image classification into semi-supervised image segmentation.

Additionally, Chapter 4 investigates the reasons behind the success of consistency regularization in semi-supervised learning for segmentation. It is discovered that consistency regularization improves model calibration. The proposed framework, MisMatch, not only outperformed previous semi-supervised segmentation methods in an in-house implementation but also surpassed these methods on the LA dataset when following established preprocessing steps using a common public codebase.

7.1.2 Chapter 5

Chapter 5 resulted in a peer-reviewed publication for MICCAI 2022, which was fortunately shortlisted for the Young Scientist Award. The extended version of that MICCAI publication is currently under review as an invited journal paper at Medical Image Analysis. Additionally, the method proposed in Chapter 5 led to a commercial patent application.

The major contribution of Chapter 5 is a new formulation of pseudo-labelling as a latent variable model, treating pseudo-labelling from a novel perspective in generative modelling within the context of binary semi-supervised segmentation. This new formulation led to the discovery of the generalisation of pseudo-labels and their learning-based approximation.

In our experiments, we evaluated the performance of pseudo-labels in semi-supervised segmentation and reported that pseudo-labelling offers significant advantages in terms of efficiency and robustness over other methods used in semi-supervised segmentation.

7.1.3 Chapter 6

Chapter 6 revisits the classical problem of parameter estimation from MRI signals, which is traditionally solved by voxel-wise fitting methods. In Chapter 6, we propose a new method to capture the global representation of all voxels in a latent space. To model the distribution in the latent space using variational inferences, two priors were tested: namely, a univariate Gaussian and a Gaussian mixture.

In our experiments, we observed that our reconstruction algorithm successfully recovers more anatomical structures than baseline methods based on single-voxel fitting. In the visualisation of the latent space, we found that the samples from the 1st and 2nd dimensions follow the distributions of the prior distributions.

7.1.4 Comparison between MisMatch and Bayesian Pseudo Labels

In the quantitative results for the segmentation of vessels in CARVE2014, Bayesian Pseudo Labels (BPL) achieved a higher intersection over union (IoU) score than

MisMatch. For instance, when both were trained on 30 labelled slices, MisMatch's IoU stood at 63.59%, whereas BPL achieved an IoU of 79%. It's worth mentioning that the base model of BPL in the CARVE experiments is a 3D U-Net, while MisMatch's base model in the CARVE experiments is a 2D U-Net.

Although the 3D U-Net base model might have some advantages over the 2D U-Net, the substantial performance improvements of BPL over CARVE suggest that BPL can achieve better segmentation accuracy than MisMatch. Additionally, from an engineering standpoint, it is easier to integrate BPL into existing segmentation models than it is to incorporate MisMatch.

7.2 Limitations

7.2.1 Chapter 4

Although MisMatch surpasses previous methods in performance, it suffers from increased model complexity. Future work should incorporate parameter sharing. For instance, the main branch could be shared across both decoders. The current implementation of MisMatch supports only binary segmentation.

7.2.2 Chapter 5

The first limitation is that once the model begins to overfit, it becomes overconfident, predicting with very high confidence, while the learned threshold also tends to oscillate around the prior mean (see an example of the learned threshold in Fig.5.14). In this situation, if the prior mean is too low, then the learned threshold will not be able to mask out the inaccurately overconfident pseudo labels. Thus, calibration becomes very important here. In future work, one could extend the formulation of pseudo labels to take calibration into account.

The second limitation is related to the use of the prior in the current paper. We use the Univariate Gaussian due to its simplicity and ease of implementation. However, a Gaussian prior might not be the most optimal choice here.

7.2.3 Chapter 6

There are limitations to our current form of the VAE algorithm that motivate future work. Due to the stochastic nature of our method, the inferred parameter maps may change, potentially hindering the repeatability and reliability of measurements. Our approach shows the most immediate benefits with more complex models (Figure 6.4), while subtler differences are seen in simpler models (Figure 6.5). We also notice a trade-off in performance: the model might excel with one parameter at the expense of others. Additionally, we assume a fixed acquisition scheme, meaning our VAE algorithm requires the exact same acquisition parameters.

7.3 Future Work

7.3.1 Chapter 4

Inspired by recent breakthroughs in foundational models, we aim to compare semi-supervised learning with few-shot fine-tuning on foundational models in segmentation tasks. To increase the impact of MisMatch, we plan to implement MisMatch on top of nnU-Net [63] or project MONAI. Additionally, a more efficient implementation is needed to reuse the main branches across the two decoders, which could also introduce extra regularisation effects. Future work should extend MisMatch to multi-class 3D tasks. Consistency in multi-class predictions could provide additional regularisation, leading to better performance. We also aim to enhance MisMatch by integrating it with existing temporal ensemble techniques [1].

7.3.2 Chapter 5

The most foreseeable future work involves integrating BPL into nnU-Net [63] or project MONAI. It could also be interesting to extend BPL to classification tasks. The current variational inference only learns a single threshold across all pixels. It would be important to further extend this towards pixel-wise thresholds with prior structural knowledge of the pixels.

The implementation of the multi-class version of BPL could be further simplified. We also aim to reuse the learned thresholds on labelled data to improve the supervised learning component. Future work could also explore the impact of other priors for the learned threshold. Potential prior distributions include categorical and Beta distributions.

Another interesting area for future work would be studying the impact of labelled data in preventing collapsed representations. Other future work could examine the convergence property of SegPL-VI and improvements in SegPL-VI for uncertainty quantification. The feasibility of applying the proposed methods to other tasks such as classification and registration also remains unexplored.

7.3.3 Chapter 6

In the future, we aim to develop algorithms that can learn a general mapping between an arbitrary acquisition sequence and model parameters, as opposed to the fixed acquisition scheme currently in use. In the current implementation with a Gaussian Mixture prior, Gumbel sampling is performed to find the index for the Gaussian component. We aim to extend this framework in the future and use Gumbel sampling to determine the appropriate number of Gaussian components. To increase the impact of the proposed model, we also aim to validate it on other, more complicated MRI models and other modalities such as hyperspectral imaging.

Appendix A

Research Paper Declaration Form (Chapter 4)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** Learning Morphological Feature Perturbations for Calibrated Semi-Supervised Segmentation

   (b) **Please include a link to or doi for the work:** `https://proceedings.mlr.press/v172/xu22a/xu22a.pdf`

   (c) **Where was the work published?** International Conference on Medical Imaging with Deep Learning

   (d) **Who published the work?** PMLR

   (e) **When was the work published?** 6th July 2022

   (f) **List the manuscript's authors in the order they appear on the publication:** Mou-Cheng Xu, Yukun Zhou, Chen Jin, Stefano B Blumberg, Frederick Wilson, Marius De Groot, Daniel C Alexander, Neil Oxtoby, Joseph Jacob

   (g) **Was the work peer reviewd?** Yes

   (h) **Have you retained the copyright?** Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

If ‘No’, please seek permission from the relevant publisher and check the box next to the below statement:

☒ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server ’e.g. medRxiv’?**
   If ’Yes’, please please give a link or doi:

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript’s authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** I conceived the idea, implemented the code, performed the experiments and wrote the draft for the manuscript; my colleague Yukun provided feedback on experiments design; all of the co-authors contributed to the writing of the manuscript.

4. **In which chapter(s) of your thesis can this material be found?** Chapter 4

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

[redacted]

Date:

7th July 2023

Supervisor/Senior Author signature (where appropriate)**:**

[redacted]

Date: 7th July 2023

Appendix B

Research Paper Declaration Form (Chapter 4)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** MisMatch: Calibrated Segmentation via Consistency on Differential Morphological Feature Perturbations with Limited Labels

   (b) **Please include a link to or doi for the work:** 10.1109/TMI.2023.3273158

   (c) **Where was the work published?** IEEE Transactions on Medical Imaging

   (d) **Who published the work?** IEEE

   (e) **When was the work published?** 8th May 2023

   (f) **List the manuscript's authors in the order they appear on the publication:** Mou-Cheng Xu, Yukun Zhou, Chen Jin, Marius De Groot, Daniel C Alexander, Neil P Oxtoby, Joseph Jacob

   (g) **Was the work peer reviewd?** Yes

   (h) **Have you retained the copyright?** No

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

© [2023] IEEE. Reprinted, with permission, from [M.C. Xu et al., “MisMatch: Calibrated Segmentation via Consistency on Differential Morphological Feature Perturbations with Limited Labels”, in IEEE Transactions on Medical Imaging, May 2023]

☒ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server ’e.g. medRxiv’?**
   If ’Yes’, please please give a link or doi:

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript’s authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** I conceived the idea, implemented the code, performed the experiments and wrote the draft for the manuscript; my colleague Yukun provided feedback on experiments design; all of the co-authors contributed to the writing of the manuscript.

4. **In which chapter(s) of your thesis can this material be found?** Chapter 4

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

[redacted]

Date:

7th July 2023

Supervisor/Senior Author signature (where appropriate)**:**

[redacted]

Date: 7th July 2023

Appendix C

Research Paper Declaration Form (Chapter 5)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** Bayesian Pseudo Labels: Expectation Maximization for Robust and Efficient Semi-Supervised Segmentation

   (b) **Please include a link to or doi for the work:** `https://doi.org/10.1007/978-3-031-16443-9_56`

   (c) **Where was the work published?** International Conference on Medical Image Computing and Computer Assisted Interventions

   (d) **Who published the work?** Springer

   (e) **When was the work published?** 16th Sep 2022

   (f) **List the manuscript's authors in the order they appear on the publication:** Mou-Cheng Xu, Yukun Zhou, Chen Jin, Marius de Groot, Daniel C Alexander, Neil P Oxtoby, Yipeng Hu, Joseph Jacob

   (g) **Was the work peer reviewd?** Yes

   (h) **Have you retained the copyright?** No

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

If ‘No’, please seek permission from the relevant publisher and check the box next to the below statement: Springer allows authors to freely use the articles in the thesis.

☒ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server ’e.g. medRxiv’?**
   If ’Yes’, please please give a link or doi:

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript’s authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** I conceived the idea, implemented the code, performed the experiments and wrote the draft of the manuscript; my colleagues Yukun, Chen and Yipeng provided feedback on experiments designs and the notations; all of the co-authors contributed to the writing of the manuscript.

4. **In which chapter(s) of your thesis can this material be found?** Chapter 5

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

[redacted]

Date:

7th July 2023

Supervisor/Senior Author signature (where appropriate)**:**

[redacted]

Date: 7th July 2023

Appendix D

Research Paper Declaration Form (Chapter 5)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?**

   (b) **Please include a link to or doi for the work:**

   (c) **Where was the work published?**

   (d) **Who published the work?**

   (e) **When was the work published?**

   (f) **List the manuscript's authors in the order they appear on the publication:**

   (g) **Was the work peer reviewd?**

   (h) **Have you retained the copyright?**

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   □ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?** Expectation Maximization Pseudo Labelling for Segmentation with Limited Annotations

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**

   If 'Yes', please please give a link or doi: `https://arxiv.org/pdf/2305.01747.pdf`

   (c) **Where is the work intended to be published?** Invited for submission to Medical Image Analysis

   (d) **List the manuscript's authors in the intended authorship order:** Mou-Cheng Xu, Yukun Zhou, Chen Jin, Marius de Groot, Daniel C Alexander, Neil P Oxtoby, Yipeng Hu, Joseph Jacob

   (e) **Stage of publication:** Under review

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** I conceived the idea, implemented the code, performed the experiments and wrote the draft of the manuscript; my colleagues Yukun, Chen and Yipeng provided feedback on experiments designs and the notations; all of the co-authors contributed to the writing of the manuscript.

4. **In which chapter(s) of your thesis can this material be found?** Chapter 5

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

[redacted]

Date:

7th July 2023

Supervisor/Senior Author signature (where appropriate)**:**

[redacted]

Date: 7th July 2023

Appendix E

Research Paper Declaration Form (Chapter 6)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?**

   (b) **Please include a link to or doi for the work:**

   (c) **Where was the work published?**

   (d) **Who published the work?**

   (e) **When was the work published?**

   (f) **List the manuscript's authors in the order they appear on the publication:**

   (g) **Was the work peer reviewd?**

   (h) **Have you retained the copyright?**

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   □ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?** Deep Variational Parameter Mapping

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**

   If 'Yes', please please give a link or doi: No

   (c) **Where is the work intended to be published?** International Conference in Medical Imaging with Deep Learning

   (d) **List the manuscript's authors in the intended authorship order:** Mou-Cheng Xu, Yukun Zhou, Tobias Goodwin-Allcock, Kimia Firoozabadi, Joseph Jacob, Daniel C. Alexander, Paddy J. Slator

   (e) **Stage of publication:** In preparation

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** I conceived the methodology, implemented the code, performed the experiments and wrote the draft of the manuscript. My colleague Paddy Slator provided inputs of MRI side and provided the original code of the baseline and data simulation. My colleague Toby provided feedback on the mathematical formulations. All of the co-authors contributed to the writing of the manuscript.

4. **In which chapter(s) of your thesis can this material be found?** Chapter 6

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate:

[redacted]

Date:

7th July 2023

Supervisor/Senior Author signature (where appropriate)**:**

[redacted]

Date: 7th July 2023

Bibliography

[1] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” *Neural Information Processing Systems (NeurIPS)*, 2017.

[2] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, “Fixmatch: Simplifying semi-supervised learning with consistency and confidence,” *Neural Information Processing Systems (NeurIPS)*, 2020.

[3] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. aFinalyson, “Semi-supervised semantic segmentation needs strong, varied perturbations,” *British Machine Vision Conference (BMVC)*, 2020.

[4] Y. Ouali, C. Hudelot, and M. Tami, “Semi-supervised semantic segmentation with cross-consistency training,” *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[5] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” *International Conference on Machine Learning (ICML)*, 2017.

[6] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” *International Conference on Learning Reprentation Workshop*, 2017.

[7] J. Scannell, A. Blanckley, H. Boldon, and B. Warrington, “Diagnosing the decline in pharmaceutical r & d efficiency,” *Nature Reviews Drug Discovery*, 2012.

[8] S. Paul, D. Mytelka, C. Dunwiddie, C. Persinger, B. Munos, S. Lindborg, and A. Schacht, “How to improve r & d productivity: the pharmaceutical industry’s grand challenge,” *Nature Reviews Drug Discovery*, 2010.

[9] NIH, “How the lungs work,” *https://www.nhlbi.nih.gov/health-topics/how-lungs-work*, 2019.

[10] W. Osler, “The principles and practice of medicine,” *Young and Pentland*, 1898.

[11] NIH, “Idiopathic pulmonary fibrosis,” *https://www.nhlbi.nih.gov/health-topics/idiopathic-pulmonary-fibrosis*, 2019.

[12] S. LA, M. EL, B. KK, and et al, “Developing disease activity and response criteria in connective tissue disease-related interstitial lung disease,” *J Rheumatol*, 1898.

[13] P. IN, J. Y, K. DS, and et al, “Clinical course and lung function change of idiopathic nonspecific interstitial pneunomia,” *European Respiratory Journal*, 2009.

[14] K. TE, A. C, B. WZ, and et al, “All-cause mortality rate in patients with idiopathic pulmonary fibrosis,” *American Journal of Respiratory and Critical Care Medicine*, 2014.

[15] V. Navaratnam, K. M. Fleming, J. West, C. J. P. Smith, R. G. Jenkins, A. Fogarty, and R. B. Hubbard, “The rising incidence of idiopathic pulmonary fibrosis in the uk,” *Throax*, 2011.

[16] NHS, “Idiopathic pulmonary fibrosis,” *https://www.nhs.uk/conditions/idiopathic-pulmonary-fibrosis/diagnosis/*, 2019.

[17] K. Flaherty, G. Toews, W. Travis, and et al, “Clinical significance of histological classification of idiopathic interstitial pneumonia,” *European Respiratory Journal*, 2002.

[18] H. Woodcock, J. Eley, D. Guillotin, and et al, “The mtorc1/4e-bp1 axis represents a critical signaling node during fibrogenesis,” *Nature Communications*, 2019.

[19] T. Hallstrand, L. Boitano, W. Johnson, C. Spada, J. Hayes, and G. Raghu, “The timed walk test as a measure of severity and survival in idiopathic pulmonary fibrosis,” *European Respiratory Journal*, 2005.

[20] H. Robbie, C. Daccord, F. Chua, and A. Devaraj, “Evaluating disease severity in idiopathic pulmonary fibrosis,” *European Respiratory Journal*, 2017.

[21] Y. Kondoh, H. Taniguchi, K. Kataoka, T. Furukawa, M. Ando, and et al, “Disease severity staging system for idiopathic pulmonary fibrosis in japan,” *Respirology*, 2017.

[22] J. Ryu, T. Moua, C. Daniels, and et al, “Idiopathic pulmonary fibrosis: Evolving concepts,” *Mayo Clinical Proceeding*, 2014.

[23] J. Egan, F. Martinez, A. Wells, and T. Williams, “Lung function estimates in idiopathic pulmonary fibrosis: the potential for a simple classification,” *Thorax*, 2004.

[24] K. Flaherty, J. Mumford, S. Murray, and et al, “Prognostic implications of physiologic and radiographic changes in idiopathic interstitial pneumonia,” *American Journal of Respiratory and critical care medicine*, 2003.

[25] S. Misumi and D. Lynch, “Idiopathic pulmonary fibrosis/usual interstitial pneumonia,” *American Journal of Respiratory and Critical Care Medicine*, 2006.

[26] S. Walsh, A. Devaraj, J. I. Enghelmayer, K. Kishi, and et al, “Role of imaging in progressive-fibrosing interstitial lung diseases,” *European Respiratory Journal*, 2018.

[27] D. Hansell, J. Goldin, T. King, D. Lynch, and et al, “Ct staging and monitoring of fibrotic interstitial lung diseases in clinical practice and treatment trials: a position paper from the fleischner society,” *The Lancet Resiratory Medicine*, 2015.

[28] H. J. Park, S. M. Lee, J. W. Song, S. M. Lee, and et al, “Texture-based automated quantitative assessment of regional patterns on initial ct in patients with idiopathic pulmonary fibrosis: Relationship to decline in forced vital capacity,” *American Journal of Roentgenology*, 2016.

[29] S. M. Humphries, J. J. Swigris, K. K. Brown, M. Strand, and et al, “Quantitative high-resolution computed tomography fibrosis score: performance characteristics in idiopathic pulmonary fibrosis,” *European Respiratory Journal*, 2018.

[30] J. Jacob, B. Brian, R. Srinivasan, K. Maria, and et al, “Automated quantitative computed tomography versus visual computed tomography scoring in idiopathic pulmonary fibrosis,” *Journal of Thoracic Imaging*, 2016.

[31] K. Oda, H. Ishimoto, K. Yatera, K. Naito, T. Ogoshi, and et al, “High-resolution ct scoring system-based grading scale predicts the clinical outcomes in patients with idiopathic pulmonary fibrosis,” *BMC Respiratory Research*, 2014.

[32] F. Maldonado, T. Moua, S. Rajagopalan, R. Karwoski, and et al, “Automated quantification of radiological patterns predicts survival in idiopathic pulmonary fibrosis,” *European Respiratory Journal*, 2014.

[33] Y. Ikuyama, A. Ushiki, M. Kosaka, J. Akahane, Y. Mukai, and et al, “Prognosis of patients with acute exacerbation of combined pulmonary fibrosis and

emphysema: a retrospective single-centre study," *BMC Pulmonary Medicine*, 2020.

[34] J. Jacob, B. Bartholmai, S. Rajagopalan, and et al, "Functional and prognostic effects when emphysema complicates idiopathic pulmonary fibrosis," *European Respiratory Journal*, 2017.

[35] A. Wells, S. Desai, M. Rubens, N. Goh, and et al, "A composite physiologic index derived from disease extent observed by computed tomography," *American Journal of Respiratory and Critical Care Medicine*, 2002.

[36] H. Robbie, A. Wells, J. Jacob, and et al, "Visual and automated ct measurements of lung volume loss in idiopathic pulmonary fibrosis," *American Journal of Roentgenology*, 2019.

[37] J. Jacob, B. Bartholmai, S. Rajagopalan, and et al, "Evaluation of computer-based computer tomography stratification against outcome models in connective tissue disease-related interstitial lung disease: a patient outcome study," *BMC Medicine*, 2016.

[38] ——, "Mortality prediction in idiopathic pulmonary fibrosis: evaluation of computer-based ct analysis with conventional severity measures," *European Respiratory Journal*, 2016.

[39] ——, "Serial automated quantitative ct analysis in idiopathic pulmonary fibrosis: functional correlations and comparison with changes in visual ct scores," *European Radiology*, 2018.

[40] ——, "Longitudinal prediction of outcome in idiopathic pulmonary fibrosis using automated ct analysis," *European Respiratory Journal*, 2018.

[41] J. Jacob, M. Pienn, C. Payer, and et al, "Quantitative ct-derived vessel metrics in idiopathic pulmonaryfibrosis: A structure–function study," *Respirology*, 2019.

[42] J. Jacob, B. Bartholmai, S. Rajagopalan, and et al, “Predicting outcomes in idiopathic pulmonary fibrosis using automated computed tomographic analysis,” *American Journal of Respiratory and Critical Care Medicine Volume*, 2018.

[43] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *CVPR*, 2015.

[44] V. Zavaletta, B. Bartholmai, and R. Robb, “High resolution multi-detector ct aided tissue analysis and quantification of lung fibrosis,” *Academic Radiology*, 2009.

[45] K. P. Murphy, *Probabilistic Machine Learning: An introduction.* MIT Press, 2022. [Online]. Available: probml.ai

[46] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Berlin, Heidelberg: Springer-Verlag, 2006.

[47] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry.* Cambridge, MA, USA: MIT Press, 1969.

[48] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[49] D. Hendrycks and K. Gimpel, “Bridging nonlinearities and stochastic regularizers with gaussian error linear units,” *CoRR*, vol. abs/1606.08415, 2016. [Online]. Available: http://arxiv.org/abs/1606.08415

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp.

5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[51] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[52] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[57] ——, “Identity mappings in deep residual networks,” *CoRR*, vol. abs/1603.05027, 2016. [Online]. Available: http://arxiv.org/abs/1603.05027

[58] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *NeurIPS*, 2012.

[60] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *MICCAI*, 2015.

[61] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, “Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation,” *Medical Image Analysis*, vol. 36, pp. 61–78, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841516301839

[62] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *MICCAI*, 2015.

[63] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. Maier-Hein, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, pp. 203 – 211, 2020.

[64] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, “Realistic evaluation of deep semi-supervised learning algorithms,” *Neural Information Processing Systems (NeurIPS)*, 2018.

[65] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, “Label propagation for deep semi-supervised learning,” *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[66] Y. Grandvalet and Y. Bengio, “Semi-supervised learning by entropy minimization,” *Neural Information Processing Systems (NeurIPS)*, 2004.

[67] D.-H. Lee, “Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks,” *ICML workshop on Challenges in Representation Learning*, 2013.

[68] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, “Semi-supervised learning with deep generative models,” *Advanced in Neural Information Processing System (NeurIPS)*, 2014.

[69] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, C. Raffel, and K. Sohn, “Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring,” *International Conference On Learning Representation (ICLR)*, 2020.

[70] J. Xu, S. Lala, B. Gagoski, E. A. Turk, P. E. Grant, P. Golland, and E. Adalsteinsson, “Semi-supervised learning for fetal brain mri quality assessment with roi consistency,” *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

[71] K. Li, S. Wang, L. Yu, and P.-A. Heng, “Dual-teacher: Integrating intra-domain and inter-domain teachers for annotation-efficient cardiac segmentation,” *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

[72] W. Cui, Y. Liu, Y. Li, M. Guo, Y. Li, X. Li, T. Wang, X. Zeng, and C. Ye, “Semi-supervised brain lesion segmentation with an adapted mean teacher model,” *Information Processing in Medical Imaging (IPMI)*, 2019.

[73] W. Hang, W. Feng, S. Liang, L. Yu, Q. Wang, K.-S. Choi, and J. Qin, “Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation,” *International conference on medical image computing and computer assisted intervention (MICCAI)*, 2020.

[74] K. Fang and W.-J. Li, “Dmnet: Difference minimization network for semi-supervised segmentation in medical images,” *International Conference on*

Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2020.

[75] X. Li, L. Yu, H. Chen, C. W. Fu, and P. A. Heng, "Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model," *British Machine Vision Conference (BMVC)*, 2018.

[76] Z. Ke, D. Qiu, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," *European Conference on Computer Vision (ECCV)*, 2020.

[77] X. Luo, J. Chen, S. Tao, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," *International Conference on Artificial Intelligence (AAAI)*, 2021.

[78] Y. Shi, J. Zhang, T. Ling, J. Lu, Y. Zheng, Q. Yu, L. Qi, and Y. Gao, "Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation," *IEEE Transactions on Medical Imaging (TMI)*, 2022.

[79] K. Chaitanya, N. Karani, C. F. Baumgartner, A. Becker, O. Donati, and E. Konukoglu, "Semi-supervised and task-driven data augmentation," *Information Processing In Medical Imaging (IPMI)*, 2019.

[80] C. Chen, C. Qin, H. Qiu, C. Ouyang, S. Wang, L. Chen, G. Tarroni, W. Bai, and D. Rueckert, "Realistic adversarial data augmentation for mr image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

[81] H. Kervadec, J. Dolz, Éric Granger, and I. B. Ayed, "Curriculum semi-supervised segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019.

[82] S. Chen, G. Bortsova, A. G.-U. Juárez, G. van Tulder, and M. de Bruijne, "Multi-task attention-based semi-supervised learning for medical image

segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019.

[83] Y. Zhu, J. Yang, S.-Q. Liu, and R. Zhang, "Inherent consistent learning for accurate semi-supervised medical image segmentation," 2023.

[84] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 87, p. 102792, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523000531

[85] P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang, and C. Desrosiers, "Cat: Constrained adversarial training for anatomically-plausible semi-supervised segmentation," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2023.

[86] M.-C. Xu, Y. Zhou, C. Jin, S. B. Blumberg, F. J. Wilson, M. de Groot, D. C. Alexander, N. P. Oxtoby, and J. Jacob, "Learning morphological feature perturbations for calibrated semi-supervised segmentation," *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2022.

[87] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between cnn and transformer," *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2022.

[88] P. Qiao, H. Li, G. Song, H. Han, Z. Gao, Y. Tian, Y. Liang, X. Li, S. K. Zhou, and J. Chen, "Semi-supervised ct lesion segmentation using uncertainty-based data pairing and swapmix," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1546–1562, 2023.

[89] Z. Wang, W. Zhao, Z. Ni, and Y. Zheng, "Adversarial vision transformer for medical image semantic segmentation with limited annotations," in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK,*

November 21-24, 2022. BMVA Press, 2022, p. 1002. [Online]. Available: https://bmvc2022.mpi-inf.mpg.de/1002/

[90] J. Chen, J. Zhang, K. Debattista, and J. Han, "Semi-supervised unpaired medical image segmentation through task-affinity consistency," *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 594–605, 2023.

[91] R. Gu, J. Zhang, G. Wang, W. Lei, T. Song, X. Zhang, K. Li, and S. Zhang, "Contrastive semi-supervised learning for domain adaptive segmentation across similar anatomical structures," *IEEE Transactions on Medical Imaging*, vol. 42, no. 1, pp. 245–256, 2023.

[92] Q. Jin, H. Cui, C. Sun, J. Zheng, L. Wei, Z. Fang, Z. Meng, and R. Su, "Semi-supervised histological image segmentation via hierarchical consistency enforcement," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 3–13.

[93] X. Zhao, Z. Wu, S. Tan, D.-J. Fan, Z. Li, X. Wan, and G. Li, "Semi-supervised spatial temporal attention network for video polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 456–466.

[94] Y. Meng, H. Zhang, Y. Zhao, D. Gao, B. Hamill, G. Patri, T. Peto, S. Madhusudhan, and Y. Zheng, "Dual consistency enabled weakly and semi-supervised optic disc and cup segmentation with dual adaptive graph convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 416–429, 2023.

[95] X. Luo, G. Wang, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, D. N. Metaxas, and S. Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Medical Image Analysis*,

vol. 80, p. 102517, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522001645

[96] W. Huang, C. Chen, Z. Xiong, Y. Zhang, X. Chen, X. Sun, and F. Wu, “Semi-supervised neuron segmentation via reinforced consistency learning,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3016–3028, 2022.

[97] Y. Lin, H. Yao, Z. Li, G. Zheng, and X. Li, “Calibrating label distribution for class-imbalanced barely-supervised knee segmentation,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 109–118.

[98] K. Wang, B. Zhan, C. Zu, X. Wu, J. Zhou, L. Zhou, and Y. Wang, “Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning,” *Medical Image Analysis*, vol. 79, p. 102447, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522000925

[99] H. Wu, J. Liu, F. Xiao, Z. Wen, L. Cheng, and J. Qin, “Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion,” *Medical Image Analysis*, vol. 78, p. 102397, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522000494

[100] L.-L. Zeng, K. Gao, D. Hu, Z. Feng, C. Hou, P. Rong, and W. Wang, “Ss-tbn: A semi-supervised tri-branch network for covid-19 screening and lesion segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 427–10 442, 2023.

[101] M.-C. Xu, Y. Zhou, C. Jin, M. de Groot, D. C. Alexander, N. P. Oxtoby, Y. Hu, and J. Jacob, “Bayesian pseudo labels: Expectation maximization for robust and efficient semi-supervised segmentation,” *International Con-*

ference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2022.

[102] F. Wu and X. Zhuang, "Minimizing estimated risks on unlabeled data: A new formulation for semi-supervised medical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6021–6036, 2023.

[103] J. Xiang, P. Qiu, and Y. Yang, "Fussnet: Fusing two sources of uncertainty for semi-supervised medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 481–491.

[104] V. Nath, D. Yang, H. R. Roth, and D. Xu, "Warm start active learning with proxy labels and selection via semi-supervised fine-tuning," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 297–308.

[105] Y. Meng, X. Chen, H. Zhang, Y. Zhao, D. Gao, B. Hamill, G. Patri, T. Peto, S. Madhusudhan, and Y. Zheng, "Shape-aware weakly/semi-supervised optic disc and cup segmentation with regional/marginal consistency," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 524–534.

[106] J. Wu, B. Shen, H. Zhang, J. Wang, Q. Pan, J. Huang, L. Guo, J. Zhao, G. Yang, X. Li, and D. Ding, "Semi-supervised learning for nerve segmentation in corneal confocal microscope photography," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 47–57.

[107] H. Basak, S. Ghosal, and R. Sarkar, “Addressing class imbalance in semi-supervised image segmentation: A study on cardiac mri,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 224–233.

[108] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin, “Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2022, pp. 11 656–11 665. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01137

[109] X. Liu, F. Xing, N. Shusharina, R. Lim, C.-C. Jay Kuo, G. El Fakhri, and J. Woo, “Act: Semi-supervised domain-adaptive medical image segmentation with asymmetric co-training,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 66–76.

[110] T. Lei, D. Zhang, X. Du, X. Wang, Y. Wan, and A. K. Nandi, “Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1265–1277, 2023.

[111] F. Lyu, M. Ye, J. F. Carlsen, K. Erleben, S. Darkner, and P. C. Yuen, “Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 797–809, 2023.

[112] D. Xiang, S. Yan, Y. Guan, M. Cai, Z. Li, H. Liu, X. Chen, and B. Tian, “Semi-supervised dual stream segmentation network for fundus lesion seg-

mentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 713–725, 2023.

[113] N. Shen, T. Xu, Z. Bian, S. Huang, F. Mu, B. Huang, Y. Xiao, and J. Li, "Scanet: A unified semi-supervised learning framework for vessel segmentation," *IEEE Transactions on Medical Imaging*, 2022.

[114] J. Wang and T. Lukasiewicz, "Rethinking bayesian deep learning methods for semi-supervised volumetric medical image segmentation," 2022.

[115] C. Chen, K. Zhou, Z. Wang, and R. Xiao, "Generative consistency for semi-supervised cerebrovascular segmentation from tof-mra," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 346–353, 2023.

[116] H. Cai, S. Li, L. Qi, Q. Yu, Y. Shi, and Y. Gao, "Orthogonal annotation benefits barely-supervised medical image segmentation," 2023.

[117] H. Basak and Z. Yin, "Pseudo-label guided contrastive learning for semi-supervised medical image segmentation," 2023.

[118] ——, "Pseudo-label guided contrastive learning for semi-supervised medical image segmentation," 2023.

[119] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, and Z. Xu, "Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 83, p. 102656, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522002845

[120] J. S. Birdle, A. J. Heading, and D. J. Mackay, "Unsupervised classifiers, mutual information and phantom targets," *Advances in Neural Information Processing Systems (NeurIPS)*, 1991.

[121] B. Athiwaratkun, M. Finzi, P. Izmailov, and A. G. Wilson, "There are many consistent explanations of unlabeled data: Why you should average," *International Conference on Learning Representations (ICLR)*, 2019.

[122] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *NeurIPS*, 2016.

[123] M.-C. Xu, N. P. Oxtoby, D. C. Alexander, and J. Jacob, "Learning to pay attention to mistakes," *British Machine Vision Conference (BMVC)*, 2020.

[124] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semisupervised semantic segmentation," *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[125] A. Veit, M. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," *Neural Information Processing Systems (NeurIPS)*, 2016.

[126] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *International Conference on 3D Vision (3DV)*, 2016.

[127] J.-P. Charbonnier, M. Brink, F. Ciompi, E. T. Scholten, C. M. Schaefer-Prokop, and E. M. van Rikxoort, "Automatic pulmonary artery-vein separation and classification in computed tomography using tree partitioning and peripheral vessel matching," *IEEE Transaction on Medical Imaging*, 2015.

[128] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, and et al, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transaction on Medical Imaging*, 2015.

[129] Z. Xiong, Q. Xia, Z. Hu, N. Huang, S. Vesal, N. Ravikumar, A. Maier, C. Li, Q. Tong, W. Si *et al.*, "A global benchmark of algorithms for segmenting late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical Image Analysis*, 2020.

[130] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self ensembleing model for semi supervised 3d left atrium segmentation," *Inter-*

national Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2019.

[131] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Schneider, B. Landman, G. Ligjens, B. Menze, O. Ronneberger, R. SUmmers, B. Ginneken, M. Bilello, P. Bilic, P. Christ, R. Do, M. Gollub, S. H. Heckers, H. Huisman, W. R. Jarnagin, M. K. McHugo, S. Napel, J. S. G. Pernicka, K. Rhode, C. Tobon-Gomez, E. Vorontsov, J. A. Meakin, S. Ourselin, M. Wiesenfarth, P. Arbeláez, B. Bae, S. Chen, L. Daza, J. Feng, B. He, F. Isensee, Y. Ji, F. Jia, I. Kim, K. Maier-Hein, D. Merhof, A. Pai, B. Park, M. Perslev, R. Rezaiifar, O. Rippel, I. Sarasua, W. Shen, J. Son, C. Wachinger, L. Wang, Y. Wang, Y. Xia, D. Xu, Z. Xu, Y. Zheng, A. L. Simpson, L. Maier-Hein, and M. J. Cardoso, "The medical segmentation decathlon," *Nature Communications*, 2022.

[132] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," *Neural Information Processing System (NeurIPS)*, 2019.

[133] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," *Neural Information Processing Systems (NeurIPS)*, 2020.

[134] M. DeGroot and S. Feinberg, "The comparison and evaluation of forecasters," *The statistician*, 1983.

[135] J. Kaplan, S. McCandlish, T. Henighan, T. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *https://arxiv.org/pdf/2001.08361.pdf*, 2020.

[136] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," *International Conference on Learning Representations (ICLR)*, 2021.

[137] M.-C. Xu, Y.-K. Zhou, C. Jin, F. Wilson, S. Blumberg, M. de Groot, D. C. Alexander, N. P. Oxtoby, and J. Jacob, "Learning morphological feature perturbations for calibrated semi supervised segmentation," *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2022.

[138] D. Berthelot, N. Carlini, I. Goodfellow, A. Oliver, N. Papernot, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Neural Information Processing Systems (NeurIPS)*, 2019.

[139] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[140] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labelling: An uncertainty-aware pseudo-label selective framework for semi-supervised learning," *International Conference on Learning Representation (ICLR)*, 2021.

[141] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," *International Conference on Computer Vision (ICCV)*, 2021.

[142] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister, "Pseudoseg: Designing pseudo labels fof semantic segmentation," *International Conference on Learning Representation (ICLR)*, 2021.

[143] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," *https://arxiv.org/abs/1908.02983*, 2019.

[144] Y. Xie, J. Zhang, Z. Liao, J. Verjans, C. Shen, and Y. Xia, “Pairwise relation learning for semi-supervised gland segmentation,” *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

[145] K. Ta, S. S. Ahn, J. C. Stendahl, A. J. Sinusas, and J. S. Duncan, “A semi-supervised joint network for simultaneous left ventricular motion tracking and segmentation in 4d echocardiography,” *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

[146] M. N. N. To, S. Sankineni, S. Xu, B. Turkbey, P. A. Pinto, V. Moreno, M. Merino, B. J. Wood, and J. T. Kwak, “Improving dense pixelwise prediction of epithelial density using unsupervised data augmentation for consistency regularization,” *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

[147] B. Unnikrishnan, C. M. Nguyen, S. Balaram, C. S. Foo, and P. Krishnaswamy, “Semi-supervised classification of diagnostic radiographs with noteacher: A teacher that is not mean,” *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

[148] H. Yang, C. Shan, A. F. Kolen, and P. H. N. de With, “Deep q-network-driven catheter segmentation in 3d us by hybrid constrained semi-supervised learning and dual-unet,” *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

[149] G. Fotedar, N. Tajbakhsh, S. Ananth, and X. Ding, “Extreme consistency: Overcoming annotation scarcity and domain shifts,” *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

[150] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert, “Semi-supervised learning for network-based cardiac mr image segmentation,” *MICCAI*, 2017.

[151] G. Wang, S. Zhai, G. Lasio, B. Zhang, B. Yi, S. Chen, T. J. Macvittie, D. Metaxas, J. Zhou, and S. Zhang, “Semi-supervised segmentation of radi-

ationinduced pulmonary fibrosis from lung ct scans with multi-scale guided dense attention," *IEEE Transactions on Medical Imaging*, 2014.

[152] Y. Wu, M. Xu, Z. Ge, J. Cai, and L. Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," *MICCAI*, 2021.

[153] O. Cahpelle, B. Scholkopf, and A. Zien, "Semi-supervised learning," *MIT Press*, 2006.

[154] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *ICLR*, 2014.

[155] ——, "Auto-Encoding Variational Bayes," Dec. 2022.

[156] Y. Li, J. Chen, X. Xie, K. Ma, and Y. Zheng, "Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation," *MICCAI*, 2020.

[157] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *International Conference on Learning Representation (ICLR)*, 2018.

[158] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Neural Information Processing System (NeurIPS)*, 2017.

[159] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[160] M. Palombo, A. Ianus, M. Guerreri, D. Nunes, D. C. Alexander, N. Shemesh, and H. Zhang, "SANDI: A compartment-based model for non-invasive apparent soma and neurite imaging by diffusion MRI," *NeuroImage*, vol. 215, pp. 116 835–116 835, Jul. 2020.

[161] N. G. Gyori, M. Palombo, C. A. Clark, H. Zhang, and D. C. Alexander, "Training data distribution significantly impacts the estimation of tissue microstructure with machine learning," *Magnetic Resonance in Medicine*, vol. 87, no. 2, pp. 932–947, Feb. 2022.

[162] S. C. Epstein, T. J. P. Bray, M. Hall-Craggs, and H. Zhang, "Choice of training label matters: How to best use deep learning for quantitative MRI parameter estimation," *ArXiv Preprint*, May 2022.

[163] S. Barbieri, O. J. Gurney-Champion, R. Klaassen, and H. C. Thoeny, "Deep learning how to fit an intravoxel incoherent motion model to diffusion-weighted MRI," *Magnetic Resonance in Medicine*, vol. 83, no. 1, pp. 312–321, Jan. 2020.

[164] M. R. Orton, D. J. Collins, D.-M. M. Koh, and M. O. Leach, "Improved intravoxel incoherent motion analysis of diffusion weighted imaging by data driven Bayesian modeling," *Magnetic Resonance in Medicine*, vol. 71, no. 1, pp. 411–420, 2014.

[165] S. D. Vasylechko, S. K. Warfield, O. Afacan, and S. Kurugol, "Self-supervised IVIM DWI parameter estimation with a physics based forward model," *Magnetic Resonance in Medicine*, vol. 00, pp. 1–11, 2021.

[166] P. J. Slator, J. Hutter, R. V. Marinescu, M. Palombo, L. H. Jackson, A. Ho, L. C. Chappell, M. Rutherford, J. V. Hajnal, and D. C. Alexander, "Data-Driven multi-Contrast spectral microstructure imaging with InSpect: INtegrated SPECTral component estimation and mapping," *Medical Image Analysis*, vol. 71, pp. 102 045–102 045, 2021.

[167] L. Manduchi, R. Marcinkevičs, M. C. Massi, T. Weikert, A. Sauter, V. Gotta, T. Müller, F. Vasella, M. C. Neidert, M. Pfister, B. Stieltjes, and J. E. Vogt, "A Deep Variational Approach to Clustering Survival Data," in *ICLR 2022*. arXiv, 2022.

[168] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *International Conference on Learning Representations*, 2017.

[169] C. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," *International Conference on Learning Representations*, 2017.

[170] R. N. Henriques, S. N. Jespersen, and N. Shemesh, "Microscopic anisotropy misestimation in spherical-mean single diffusion encoding MRI," *Magnetic Resonance in Medicine*, vol. 81, no. 5, pp. 3245–3261, 2019.

[171] T. Behrens, M. Woolrich, M. Jenkinson, H. Johansen-Berg, R. Nunes, S. Clare, P. Matthews, J. Brady, and S. Smith, "Characterization and propagation of uncertainty in diffusion-weighted MR imaging," *Magnetic Resonance in Medicine*, vol. 50, no. 5, pp. 1077–1088, Nov. 2003.

[172] E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. van der Walt, M. Descoteaux, I. Nimmo-Smith, and Dipy Contributors, "Dipy, a library for the analysis of diffusion MRI data," *Frontiers in Neuroinformatics*, vol. 8, Feb. 2014.

[173] R. H. Fick, D. Wassermann, and R. Deriche, "The Dmipy Toolbox: Diffusion MRI Multi-Compartment Modeling and Microstructure Recovery Made Easy," *Frontiers in Neuroinformatics*, vol. 13, no. October, pp. 1–26, 2019.

[174] J. P. Lim, S. B. Blumberg, N. Narayan, S. C. Epstein, D. C. Alexander, M. Palombo, and P. J. Slator, "Fitting a Directional Microstructure Model to Diffusion-Relaxation MRI Data with Self-Supervised Machine Learning," Oct. 2022.

[175] D. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. Curtiss, S. Della Penna, D. Feinberg, M. Glasser, N. Harel, A. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. Petersen, F. Prior, B. Schlaggar, S. Smith,

A. Snyder, J. Xu, and E. Yacoub, “The Human Connectome Project: A data acquisition perspective,” *NeuroImage*, vol. 62, no. 4, pp. 2222–2231, Oct. 2012.

[176] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson, “The minimal preprocessing pipelines for the Human Connectome Project,” *NeuroImage*, vol. 80, pp. 105–124, Oct. 2013.