

Imperial College of Science, Technology and Medicine
Department of Computing

Machine Learning for Magnetic Resonance Image Reconstruction and Analysis

Chen Qin

A dissertation submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy
of
Imperial College London

December, 2019

Abstract

Machine learning has great potentials to improve the entire medical imaging pipeline, providing support for clinical decision making and computer-aided diagnosis. This thesis mainly focuses on developing machine learning methods for the improvement of magnetic resonance (MR) image reconstruction and analysis, specifically on dynamic MR image reconstruction, image registration and segmentation.

Firstly, we propose to tackle the white matter hyperintensity (WMH) segmentation problem present in elderly subjects or patients with vascular diseases. A two-step framework consisting of (semi-)supervised large margin based algorithms is proposed, where both common features shared across subjects and individual-specific information from target subject are considered and utilised. To further improve the segmentation and differentiate WMHs and stroke lesions, a deep learning based model, uResNet, is proposed. It combines the effective U-net architecture with residual elements, and considers a sampling strategy for imbalanced data. Experiments demonstrate its better performance than other competing methods.

Secondly, we propose to address the problem of dynamic MRI reconstruction from highly undersampled k -space data. A convolutional recurrent neural network architecture (CRNN-MRI) is developed, where it embeds the iterative optimisation process in a learning setting and exploits the temporal redundancies of cardiac sequences. As a complement, a k - t NEXT network is presented to recover the signals in both x - f and image domains alternately. We show that our proposed models can effectively reconstruct high quality cardiac MRI at high acceleration factors.

Finally, deep learning approaches for image registration and its applications are investigated. A joint learning framework for cardiac segmentation and motion estimation is proposed, which can provide multi-task predictions simultaneously. This motion estimation method is further extended for multi-modal deformable registration, in which it proposes to embed multi-modal images onto a common latent shape domain via disentangled representations. Experimental results indicate their competing performance and faster registration speed compared to other conventional methods.

©The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives license. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the license terms of this work.

Declaration of Originality

I declare that the work presented in this thesis is my own, unless specifically acknowledged.

Chen Qin

Acknowledgements

I would like to thank my esteemed supervisor, Prof. Daniel Rueckert, for offering me the opportunity to pursue a PhD degree and work under his supervision at the biomedical image analysis group (BioMedIA). I am very grateful for his invaluable guidance and constant support during my PhD study, which have greatly encouraged me to solve difficult research problems and inspired me with new ideas for my research.

I would also like to thank all my colleagues at the BioMedIA group for their great helps. They have taught me and supported me a lot over the last four years, and have brought many fun and memorable moments to the lab. In addition, I really appreciate all the co-authors and collaborators, who helped my work to become better and get published in top-tier journals and conferences. Special thanks to Dr. Wenjia Bai and Jo Schlemper for their helpful suggestions and kind encouragements, as well as their constant supports and contributions to my research. Also great thanks to my examiners Dr. Ender Konukoglu and Prof. Ferdinando Rodriguez y Baena for the insightful discussions and valuable advice to improve the thesis better.

Besides, I would also like to express my sincere gratitude to my beloved family for their endless love. I am particularly thankful to my parents who supported me to study at Imperial College London, and have always given me the courage and strength to pursue my goals.

Acronyms

ANN Artificial Neural Network

BP Backpropagation

BraTS brain tumor segmentation

CNN Convolutional Neural Network

COPD Chronic Obstructive Pulmonary Disease

CSF Cerebrospinal Fluid

CT Computed Tomography

DNN Deep Neural Network

ED End-Diastolic

EF Ejection Fraction

ES End-Systolic

FCN Fully Convolutional Network

FFD Free Form Deformation

FLAIR Fluid-attenuated Inversion Recovery

FN False Negative

FP False Positive

GAN Generative Adversarial Network

GM Grey Matter

HD Hausdoff Distance

LSVRC Large Scale Visual Recognition Challenge

LV Left Ventricle

MCD Mean Contour Distance

MI Mutual Information

ML Machine Learning

MR Magnetic Resonance

MRI Magnetic Resonance Imaging

PSNR Peak-to-Noise Ratio

ReLU Rectified Linear Unit

RNN Recurrent Neural Network

ROI Region of Interest

SSD Sum of Squared Differences

SSFP Steady-State Free Precession

SSIM Structural Similarity Index Measure

STN Spatial Transformer Network

SVM Support Vector Machine

TP True Positive

WM White Matter

WMH White Matter Hyperintensity

Contents

Abstract	i
Acknowledgements	vii
Acronyms	ix
1 Introduction	1
1.1 Medical Imaging Overview	1
1.2 Contributions	5
1.3 Thesis Overview	8
1.4 Publications	9
2 Background	13
2.1 Introduction	13
2.2 Classical Machine Learning Methods	14
2.2.1 Naive Bayes Classifier	15
2.2.2 Support Vector Machines	16
2.2.3 Random Forest	17

2.2.4	Artificial Neural Network	19
2.3	Deep Learning Methods	20
2.3.1	Convolutional Neural Networks	21
2.3.2	Recurrent Neural Networks	25
2.3.3	Generative Adversarial Networks	29
2.3.4	Spatial Transformer Network	33
2.4	Conclusions	34
I	Brain Lesion Segmentation	36
3	Background: White Matter Hyperintensity (WMH) Segmentation	37
3.1	Introduction	37
3.2	General preprocessing steps	40
3.3	Related Work	41
3.3.1	Traditional Methods	41
3.3.2	Deep Learning Methods	48
3.4	Evaluation Metrics	52
3.5	Conclusions	54
4	Large Margin Algorithm for WMH Segmentation	55
4.1	Introduction	55
4.2	Background and Notation	57

4.3	Supervised and Semi-supervised Large Margin Algorithms	59
4.3.1	Supervised Large Margin Algorithm	59
4.3.2	Semi-supervised Large Margin Algorithm	61
4.4	Experiments	65
4.4.1	Data	65
4.4.2	MR image pre-processing	65
4.4.3	Experimental settings	66
4.5	Results	67
4.5.1	Parameter analysis	67
4.5.2	Experimental results on WMH segmentation	69
4.6	Conclusions	72
5	CNNs for differential segmentation of WMHs and stroke lesions	74
5.1	Introduction	74
5.2	Methods	76
5.2.1	Network architecture	77
5.2.2	Loss function and class imbalance	78
5.2.3	Data sampling and class imbalance	80
5.3	Data	82
5.3.1	MRI acquisition	82
5.4	Experiments and results	84
5.4.1	Model training	85

5.4.2	Comparison to state-of-the-art	88
5.4.3	Variations of uResNet	93
5.4.4	Additional DeepMedic experiments	94
5.4.5	Clinical evaluation	95
5.5	Discussion and Conclusions	99
II	Dynamic MR Image Reconstruction	101
6	Background: Dynamic MR Image Reconstruction	102
6.1	Introduction	102
6.2	Classical Methods for Dynamic MR Reconstruction	103
6.3	Deep Learning Methods for MR Reconstruction	107
6.3.1	Deep Learning Methods in Image Space	107
6.3.2	Deep Learning Methods in k -space	109
6.4	Conclusions	111
7	Convolutional Recurrent Neural Network for Dynamic MRI Reconstruction	112
7.1	Introduction	112
7.2	Method	114
7.2.1	Problem Formulation	114
7.2.2	CRNN for MRI reconstruction	115
7.2.3	Network Learning	120

7.3	Experiments	121
7.3.1	Dataset and Implementation Details	121
7.3.2	Evaluation Method	123
7.3.3	Results	124
7.3.4	Variations of Architecture	127
7.3.5	Feature Map Analysis	129
7.4	Discussion	130
7.5	Conclusions	131
8	Dynamic MR Image Reconstruction Exploiting k-t Correlations	133
8.1	Introduction	133
8.2	k - t NEXT for Dynamic MRI Reconstruction	134
8.2.1	xf -CNN in x - f domain	135
8.2.2	k - t NEXT in complementary domains	137
8.3	Experiments and Results	138
8.3.1	Dataset and Implementation Details	138
8.3.2	Results	138
8.4	Conclusions	140
III	Image Registration and Motion Estimation	142
9	Background: Image Registration	143
9.1	Introduction	143

9.2	Non-rigid Registration	144
9.2.1	Similarity Measures	145
9.2.2	Techniques	147
9.3	Related Work	151
9.3.1	Deformable Mono-modal Registration	152
9.3.2	Deformable Multi-modal Registration	153
9.4	Conclusions	155
10	Joint Motion Estimation and Segmentation of Cardiac MRI	156
10.1	Introduction	156
10.2	Joint Learning of Motion Estimation and Segmentation	158
10.2.1	Methods	158
10.2.2	Experiments and Results	161
10.3	Joint Motion Estimation and Segmentation from Undersampled Cardiac MRI . .	164
10.3.1	Methods	164
10.3.2	Experiments and Results	168
10.4	Conclusions	170
11	Unsupervised Multi-modal Deformable Image Registration	171
11.1	Introduction	171
11.2	Proposed Method	173
11.2.1	Image disentangling via unpaired image-to-image translation	173

11.2.2 Multi-modal image registration via disentangled representations	175
11.2.3 Multi-modal image registration via combined metrics	176
11.3 Experiments	178
11.3.1 Datasets	178
11.3.2 Experimental Settings	179
11.3.3 Results	181
11.4 Conclusions	183
12 Conclusions	185
12.1 Summary of Thesis Achievements	185
12.2 Limitations and Future Work	188
12.2.1 2D Networks vs 3D Networks	188
12.2.2 Model Generalisation	189
12.2.3 Learning with Limited Supervision	191
12.2.4 Accelerated Dynamic Parallel Imaging	193
12.2.5 Class Imbalance	193
12.2.6 Model Interpretability	194
Bibliography	194

List of Tables

1.1	Most common MRI sequences and comparisons of their TR/TE time.	3
3.1	Metrics used to evaluate segmentation performance.	52
4.1	Analysis of parameters γ_1 and γ_2 on validation set. Reported numbers are DSC score.	67
4.2	Performance comparison results on brain MR scans with WMH (The best results are shown in bold).	70
4.3	Correlation analysis with clinical variables (CC Fazekas) and correlation analysis between expert and automatic volumes (R^2 and Trend) (The best results are shown in bold)	73
5.1	MR imaging sequence details for the three acquisition protocols used.	83
5.2	Mean Dice scores of WMH and stroke, for different inputs with uResNet and DeepMedic. Difference in Dice score between the two methods is given in italics. F: FLAIR image, CS: cerebro-spinal track atlas, WM: white matter probability map, T1: T1 weighted image.	92

5.3	Mean Dice scores of WMH and stroke (standard deviation in parenthesis), correlation analysis between expert and automatic volumes (R^2 and trend), and correlation with clinical variables. No statistical significance between uResNet and uResNet2 was observed ($p > 0.05$), while there was a statistically significant difference ($p < 0.001$) between patch off-center sampling (uResNet) and regular no off-center sampling (uResNet_NoC).	94
5.4	Mean Dice scores of WMH and stroke (standard deviation in parenthesis), correlation analysis between expert and automatic volumes (R^2 and trend), and correlation with clinical variables.	97
5.5	p-values of linear regression associations between volumes calculated with different methods and risk factors. Bold numbers indicate statistical significance above 0.05.	97
5.6	p-values of linear regression associations between volumes calculated with different methods and risk factors after residual outliers were removed. Bold numbers indicate statistical significance above 0.05.	98
7.1	Performance comparisons (MSE, PSNR:dB, SSIM, and HFEN) on dynamic cardiac data with different acceleration rates. MSE is scaled to 10^{-3} . The bold numbers are better results of the proposed methods than that of the other methods.	124
7.2	Performance comparisons on investigating the effects of each recurrence in the module. Reported results are the mean PSNR on data with undersampling factor 9 via 3-fold cross-validation. For this study, the number of iteration was set as 2.	128
7.3	Performance comparisons with different model architectures. Reported results are the mean PSNR on data with undersampling factor 9 via 3-fold cross-validation. (FPT: forward pass time; BPT: backward pass time)	128

8.1	Comparison results of different methods on dynamic cardiac cine MRI with high undersampling rate 9 and 12. Best results are indicated in bold.	140
10.1	Evaluation of segmentation accuracy for the proposed joint model and the baseline (Seg only) method in terms of the Dice Metric (mean and standard deviation).	162
10.2	Evaluation of motion estimation accuracy for FFD, proposed model for Motion only and the proposed joint model in terms of the mean contour distance (MCD) and Hausdorff distance (HD) in mm (mean and standard deviation). Time reported is testing time on 50 frames in a cardiac cycle per slice.	163
10.3	Evaluation of motion estimation accuracy for undersampled MR data with different acceleration factors in terms of the mean contour distance (MCD) and Hausdorff distance (HD) in mm (mean and standard deviation). Loss function using $\mathcal{L}_m$ (Eq. 10.3) is termed as Proposed-A, and the one using $\mathcal{L}_m$ (Eq. 10.4) is termed as Proposed-B. Bold numbers indicate the best results for different undersampling rates.	169
10.4	Evaluation of segmentation performance under different acceleration factors in terms of Dice Metric (mean and standard deviation) and average percentage (%) error for ejection fraction (EF) compared with fully-sampled data.	170
11.1	Evaluation of multi-modal registration on COPDGene dataset in terms of average Dice score, MCD, HD (unit: pixel), run time (GPU/CPU) and Grad Det-Jac ($\times 10^{-2}$) of each deformation field.	181
11.2	Evaluation of bi-directional multi-modal registration on BraTS dataset in terms of RMSE(Δ) (unit: pixel) and Grad Det-Jac ($\times 10^{-2}$) for $T2 \rightarrow T1$ registration, and RMSE(I)) for the inverse direction. Average run time (GPU/CPU) is also provided.	182

List of Figures

1.1	Visualisation comparisons of T1, T2 and FLAIR of an adult brain. Image adopted from https://casemed.case.edu/clerkships/neurology/	3
1.2	Breakdown of papers published on deep learning for medical image analysis over the years. Left: the distribution of publications using different neural network models. Right: the distribution of these publications on various medical imaging tasks. Image adopted from Litjens et al. [156].	5
2.1	Maximum-margin hyperplane and margins for an SVM trained with samples from two classes in 2D feature space. Image adopted from [56].	16
2.2	An illustrative diagram showing an example of random forest. A random forest composes of a number of decision trees. The final decision of the random forest is made by the majority voting from its decision trees. Image adopted from [69].	18
2.3	An illustrative diagram of an artificial Perceptron. The output of the Perceptron is a weighted sum of the inputs followed by a linear threshold function. Image taken from [229].	20
2.4	An illustrative diagram of a multi-layer Perceptron, with an input layer, two hidden layers and an output layer. Image adopted from http://cs231n.github.io	20
2.5	The network architecture of LeNet-5 for handwritten digits recognition. Image taken from [149].	23

2.6	The network architecture of AlexNet. Image taken from [143].	23
2.7	The network architecture of VGG-16. Image adopted from [228].	24
2.8	The architecture of residual block in ResNet. Image adopted from [107].	24
2.9	The network architecture of DenseNet. Image adopted from [116].	24
2.10	The network architecture of FCN for image semantic segmentation. Image taken from [160].	25
2.11	Diagram of an unrolled recurrent neural networks. Image adopted from [148]. . .	26
2.12	Illustration diagram of LSTM and GRU unit. Images taken from https://colah.github.io/ . . .	27
2.13	The framework of Generative Adversarial Networks. Image adopted from O'Reilly (https://www.oreilly.com).	30
2.14	The architecture of a spatial transformer module. The input feature map U is passed to a localisation network to regress the transformation parameter θ . The regular grid G over V is transformed to the sampling grid $T_\theta(G)$, which is applied to U to generate the warped feature map V . Image taken from [123].	34
3.1	An example of T1 image, FLAIR image, WM tissue, grey matter (GM) tissue and ventricle. These images are from corresponding slice of the same subject. Tissues were segmented by an automated brain segmentation tool called MALP-EM [150]. The red arrows point to WMH on the FLAIR.	38
3.2	The flow diagram of the lesion segmentation algorithm. Image adopted from [218].	42
3.3	Example segmentation results comparing LGA and LPA on subjects with NDs. .	47
3.4	Network architecture of DeepMedic for brain tumor segmentation. Image adopted from [131].	50
3.5	Network architecture of U-net for biomedical image segmentation. Image taken from [206].	51

3.6	Schematic illustration of measuring the segmentation errors for calculating the Dice similarity coefficient (DSC). Image adopted from [174].	53
4.1	The lesion atlas constructed from a different group of subjects.	64
4.2	The model's performance influenced by the value of m over the validation set. .	68
4.3	SSLM's performance influenced by different initialisation strategies over the validation set.	69
4.4	Example WMH segmentation results on subject with medium lesion load compared with LGA, LPA, UOCL, SVM and human expert annotations. (a) Original FLAIR image (b) Expert segmentation (c) LGA (d) LPA (e) UOCL (f) SVM (g) SLM (h) CS ³ LM	69
4.5	Example WMH segmentation results on subject with high lesion load compared with LGA, LPA, UOCL, SVM and human expert annotations. (a) Original FLAIR image (b) Expert segmentation (c) LGA (d) LPA (e) UOCL (f) SVM (g) SLM (h) CS ³ LM	70
4.6	Bland and Altman plot: differences between expert annotated segmentation volumes and automatic segmentation volumes vs. the mean of the two measurements, with mean difference (solid line) and the representation of the limits of agreement (dotted line), from -1.96SD to +1.96SD.	72
5.1	Proposed u-shaped residual network (uResNet) architecture for WMH segmentation and differentiation.	76
5.2	Training patch sampling.	81
5.3	Effect of different loss functions on uResNet trained using FLAIR images.	86
5.4	Different input channel exploration. F: FLAIR image, CS: cerebro-spinal track atlas, WM: white matter probability map, T1: T1 weighted image.	87

5.5	Automated versus expertly generated WMH volumes (as ICV %). The solid line indicates the linear trend $f(x)$ of the comparison, while the dotted line indicates the ideal trend $f(x) = 1.0x + 0.0$	89
5.6	Bland-Altman plots comparing expert annotations with all other methods in WMH segmentation.	90
5.7	Automated versus expertly generated stroke volumes. LPA and LGA are unable to distinguish between WMH and stroke, hence cannot generate these results. The solid line indicates the linear trend $f(x)$ of the comparison, while the dotted line indicates the ideal trend $f(x) = 1.0x + 0.0$	92
5.8	Visual comparisons of all competing methods. Yellow lines delineate WMHs, green lines stroke and white arrows point to interesting result areas. Best seen in color.	96
6.1	k-t FOCUSS reconstruction flow chart. Image adopted from [129].	106
6.2	Illustration of DC-CNN architecture. Image taken from [214].	108
6.3	Illustration of DC-CNN architecture for dynamic MR reconstruction. Image adopted from [215].	109
6.4	Schematic illustration of data sharing layer. The acquired k -space lines are color encoded. Missing entries in each frame are aggregated using values from n_{adj} neighboring frames. Overlapped lines are averaged. Image adopted from [215].	110
7.1	(a) Traditional optimisation algorithm using variable splitting and alternate minimisation approach, (b) the optimisation unrolled into a deep convolutional network incorporating the data consistency step, and (c) the proposed architecture which models optimisation recurrence.	115

- 7.2 (a) The overall architecture of proposed CRNN-MRI network for MRI reconstruction. (b) The structure of the proposed network when unfolded over iterations, in which $\mathbf{x}_{rec}^{(0)} = \mathbf{x}_u$. (c) The structure of BCRNN-t-i layer when unfolded over the time sequence. The green arrows indicate feed-forward convolutions which are denoted by W_l . The blue arrows (W_i) and red arrows (W_t) indicate recurrent convolutions over iterations and the time sequence respectively. For simplicity, we use a single notation to denote weights for these convolutions at different layers. However, in the implementation, the weights are independent across layers. 116
- 7.3 Mean PSNR values (Proposed-B) vary with the number of iterations at test time on data with different acceleration factors. Here AF stands for acceleration factor. 125
- 7.4 Visualisation results of intermediate steps during the iterations of a reconstruction. (a) Undersampled image by acceleration factor 9 (b) Ground Truth (c-l) Results from intermediate steps 1 to 10 in a reconstruction process. 126
- 7.5 The comparison of reconstructions on spatial dimension with their error maps. (a) Ground Truth (b) Undersampled image by acceleration factor 9 (c,d) Proposed-B (e,f) 3D CNN (g,h) 3D CNN-S (i,j) k-t FOCUSS (k,l) k-t SLR 126
- 7.6 The comparison of reconstructions along temporal dimension with their error maps. (a) Ground Truth (b) Undersampled image by acceleration factor 9 (c,d) Proposed-B (e,f) 3D CNN (g,h) 3D CNN-S (i,j) k-t FOCUSS (k,l) k-t SLR . . . 127
- 7.7 Cosine distances for the feature maps extracted from i th-layer of the subnetworks across 10 cascades/iterations. Top row shows $i = 1$, which corresponds to BCRNN-t-i unit for CRNN, 1st convolution layers for 3D-CNN and 3D-CNN-S. Bottom row shows $i = 4$, which corresponds to the third CRNN-i unit for CRNN, 4th convolution layers for 3D-CNN and 3D-CNN-S. In general, the distribution of $\cos(\theta)$ is closer to 0 for CRNN than for the CNN's. 129

7.8	Examples of the feature maps from the CRNN-MRI (Proposed-A), 3D CNN and 3D CNN-S, at iteration 10.	129
8.1	The k - t NEXT reconstruction diagram. True signals can be recovered by iteratively updating the reconstruction in both (a) x - f and (b) image domains via learning the xf -CNN and CRNN jointly. For mathematical notations, please refer to Eq. 8.2.	136
8.2	Detailed network architecture of xf -CNN and CRNN in k - t NEXT. Numbers inside CNN and BCRNN layer are kernel size, dilation factor, numbers of filters, and the negative slope for LeakyReLU is 0.01. DC stands for data consistency. .	139
8.3	Comparison results on spatial and temporal dimensions with their error maps. A dynamic video is shown in supplementary materials for better visualisation. .	141
8.4	Visualisation in x - f domain. (a) Ground Truth (b) k - t sampling pattern (c) $9\times$ undersampled data (d) Reconstructed x - f image (e) Error between (c) and (d). .	141
10.1	The overall schematic architecture of proposed network for joint estimation of cardiac motion and segmentation. (a) The proposed Siamese style multi-scale recurrent motion estimation branch. (b) The segmentation branch which shares the joint feature encoder with motion estimation branch. The architecture for feature encoder is adopted from VGG-16 net before FC layer. Both branches have the same head architecture as the one proposed in [21], and the concatenation layers of motion estimation branch are from last layers at different scales of the feature encoder.	158
10.2	Visualisation results for simultaneous prediction of motion estimation and segmentation. Myocardial motions are from ED to other time points. Images are cropped to the region of interest for better visualisation.	163

10.3	(a) (b) Labeling results obtained by warping the ED frame segmentation to other time points using FFD and the proposed joint model. Results are shown in temporal views of the red short-axis line. (c) Left ventricular volume (ml) of the subject by warping the ED frame segmentation to other time points in a cardiac cycle.	164
10.4	The overall schematic architecture of proposed network for joint estimation of cardiac motion and segmentation directly from undersampled data. (a) (b) The Motion-Seg net adopted from [194]. (c) Proposed architecture for training the Motion-Seg net on undersampled data. US: undersampled, FS: fully-sampled. . .	165
10.5	Comparison visualisation results for simultaneous prediction of motion estimation and segmentation on data with undersampling rates 8. Myocardial motions are from ED to other time points (numbers on the top right). Segmentations are overlaid on fully-sampled data. Images are cropped to the region of interest for better visualisation.	170
11.1	Image-to-image translation framework (Section 11.2.1). x and y are sample images in $\mathcal{X}$ and $\mathcal{Y}$ domain respectively. $\{E_{\mathcal{X}}^s, E_{\mathcal{Y}}^s\}$ and $\{E_{\mathcal{X}}^a, E_{\mathcal{Y}}^a\}$ are corresponding shape encoders and appearance encoders. $\{G_{\mathcal{X}}, G_{\mathcal{Y}}\}$ are image generators. (a) Image self-reconstruction with $\{\tilde{x}, \tilde{y}\}$ reconstructed from $\{x, y\}$. (b) Image-to-image translation and cross-cycle reconstruction. $\{u, v\}$ are translated images from $\{x, y\}$ to domain $\{\mathcal{Y}, \mathcal{X}\}$ respectively. $\{\hat{x}, \hat{y}\}$ are cross-cycle reconstructed images. Black dotted lines and tilde notation indicate the consistency between variables.	173

11.2 Overview architecture of the proposed models. (a) Multi-modal image registration via disentangled representations (Section 11.2.2). x' and y' are warped images from x and y . (b) Learning-based similarity metric in image space. u , v and $\hat{x}$, $\hat{y}$ are translated and reconstructed images respectively adopted from Fig.11.1. (a)+(b) Multi-modal image registration via combined metrics (Section 11.2.3). G_r is the registration network in latent space. D_r^x, D_r^y are discriminators in image space.	175
11.3 An illustrative architecture of the registration network in UMDIR method. . . .	177
11.4 Visualisation results of our model against baseline methods, where warped moving images, corresponding estimated deformation fields and Jacobian determinant are shown. Left: COPDGene data; Right: BraTS data with GT overlaid on Fixed image.	181
11.5 Comparison between variants of proposed methods for T1 to T2 registration. Error maps are pixel-wise intensity differences between warped images and target moving images. Red arrows indicate regions where UMDIR-Lat outperforms UMDIR-GAN, and yellow arrows point out regions where UMDIR-GAN outperforms UMDIR-Lat. Combined metrics shows better performance than each of them separately.	183

Chapter 1

Introduction

1.1 Medical Imaging Overview

Medical imaging is the set of imaging techniques that can produce visual representations of internal structures of the human body for diagnosis or treatment purposes. It plays a crucial role in improving public health for all populations. Medical imaging first started in 1895, after the discovery of x-rays, and intrigued a growing interest in radiology. Over the first half of 20th century, a lot of improvements and inventions were proposed to create effective imaging services. In 1960s, ultrasound scanning became the next major imaging technique with the development of sonar. It uses high frequency, low wavelength sound waves to penetrate through the body, and images are then produced from the varying degree of reflection by tissues. The next big strides of medical imaging came in 1970s, when tomographic imaging techniques, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), were invented and became widely available. This technique brought a lot of benefits and big changes to radiology. It made the imagery of every part of the body with enhanced quality become possible, and enabled tomographic imaging instead of projections, thus paving the way for 3D imaging.

With these imaging techniques becoming more common and more advanced, ways of analysing medical images are increasingly needed to fully exploit the contained information. Computer-based analysis gained its popularity since the emergence of digital imaging techniques, and is

now becoming gradually essential for clinical use. In clinical studies, image analysis is often used to detect patterns between multiple images and to aid clinical decision making processes such as disease diagnosis and treatment planning. It is also very important for the application of computer-assisted interventions such as surgical navigation and robotic surgery.

This thesis will mainly focus on investigating machine learning methods to improve the medical imaging pipeline, from image reconstruction to image analysis. In particular, we present the work on reconstruction, segmentation and registration of MR images. In the following, a brief overview of MRI and machine learning in medical imaging will be presented.

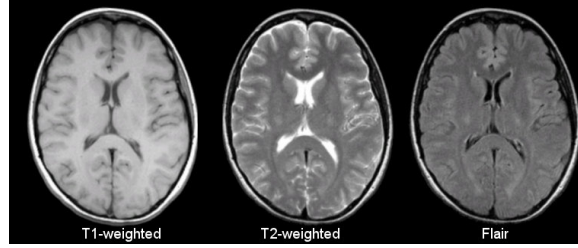
Magnetic Resonance Imaging

There are many types of medical imaging techniques, and different imaging modalities can show distinct structures or tissue features, which are used for different purposes. Among various imaging modalities, MRI is used to provide detailed visualisation of anatomy with high quality and good soft tissue contrast. In contrast to CT, MRI uses magnetic fields and radio waves to generate images, and thus there is no radiation exposure and is relatively safe.

MRI is enabled by the physical phenomenon of nuclear magnetic resonance (NMR), where nuclei in a powerful static magnetic field can be perturbed by the application of a radio frequency (RF) electromagnetic pulse, and respond by emitting a measurable RF signal [147]. When under the influence of a strong magnetic field, the hydrogen atoms will align in their spin axis. A RF pulse sequence then can be applied to cause the perturbation of this spin alignment, and the RF energy is absorbed by the atoms. Once the RF pulse is removed, the atoms lose their magnetisation through various relaxation processes, and the emitted RF energy can be captured by a MRI scanner. A Fourier transformation is then used to convert the frequency information contained in the measured emitted signals to reconstruct the image. The MR image quality is normally determined by the strength of the static magnetic field, and the spatial localisation within the body is achieved via the encoding gradients which lead to the spatial characterisation of the signals [257].

Table 1.1: Most common MRI sequences and comparisons of their TR/TE time.

	TR	TE
T1-weighted	short	short
T2-weighted	long	long
FLAIR	very long	very long

Figure 1.1: Visualisation comparisons of T1, T2 and FLAIR of an adult brain. Image adopted from <https://casemed.case.edu/clerkships/neurology/>.

Different tissues can be characterised by two different relaxation time: longitudinal relaxation time (T1) which is a measure of rate at which excited hydrogen atoms return to equilibrium (spinning atoms realign with external magnetic field), and transverse relaxation time (T2) which is related to the time taken for excited atoms to lose phase coherence. Different tissues have different T1 and T2 time constants, thus generating different signals and showing contrasts in MR images. In addition, by varying the sequence of RF pulses, i.e., the time between successive pulse sequences (TR: repetition time) and time between the pulse and echo signal (TE: echo time), different types of images can be generated. The most commonly used MRI sequences are T1-weighted, T2-weighted and Fluid Attenuated Inversion Recovery (FLAIR). A comparison of TR and TE time for these three modalities are presented in Table 1.1, and a visualisation comparisons of them are given in Fig. 1.1, showing different tissue contrast and brightness in different MRI sequences.

Machine Learning in Medical Imaging

Machine learning is an active field of research which focuses on the study of algorithms and statistical models to extract and discover meaningful patterns from examples. In recent years, machine learning algorithms have been widely applied and researched in the field of medical

image analysis [79]. In contrast to human observers, computers can perform repetitive tasks consistently and tirelessly, and the application of machine learning algorithms have also shown its capability to learn complex tasks, even beyond human perception. This has led to substantial interest in machine learning for medical image analysis, which could potentially provide support for computer-aided diagnosis and clinical decision making.

In general speaking, machine learning algorithms often begin with the image feature extraction, which is to select informative and non-redundant variables or representations that can best describe the input data with sufficient accuracy. Then a classifier is trained based on the extracted features to learn from the training data and make a prediction. Deep learning, as a branch of machine learning, is now being widely used. In contrast to traditional machine learning methods which requires feature extraction as a first step, deep learning models identify features as part of the learning process, that is to learn to detect useful representations automatically for a specific task.

Machine learning for medical imaging is an important growth area. These algorithms have been applied for many challenging tasks, such as brain tumor segmentation [131, 283], MR image reconstruction [101, 215], image registration [65, 142], and lung nodule classification [224]. There has been an increasing number of works investigating the application of deep learning for different medical image analysis tasks, and an overview of the number of publications over the years is presented in Fig. 1.2 [156]. It can be seen that most of these works focused on employing convolutional neural networks (CNNs) for medical image segmentation tasks.

Machine learning has already been applied in the practice of radiology, and computer-aided diagnosis and detection performed by these machine learning or deep learning algorithms can effectively aid radiologists interpret medical images and reduce interpretation times [79]. These applications will probably continue to grow at a rapid pace and have a greater influence in the near future. It is of great importance to engage in this area of research to better understand the machine learning algorithms for their safe and effective applications in medical imaging.

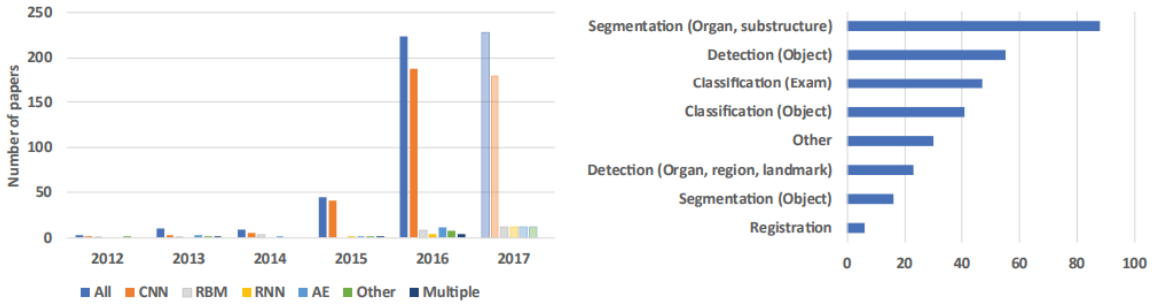


Figure 1.2: Breakdown of papers published on deep learning for medical image analysis over the years. Left: the distribution of publications using different neural network models. Right: the distribution of these publications on various medical imaging tasks. Image adopted from Litjens et al. [156].

1.2 Contributions

Machine learning has great potentials in improving the entire medical imaging pipeline, including image acquisition, reconstruction, analysis, detection and diagnosis. Among these, MR image reconstruction and analysis are two key components in the pipeline. It is known that the acquisition speed of MRI is fundamentally limited, and there has been many works investigating on the accelerated MR imaging. Most of these approaches consider undersampling the data in k -space and then reconstruct the image. However, it still remains a challenge on how to effectively exploit sparsity or spatio-temporal redundancies to reconstruct MR images at a fast reconstruction speed while achieving high quality. On the other hand, MR image analysis paves the way for computer-aided diagnosis and interventions. It provides either quantitative measurements or qualitative visualisations for disease assessment or predictions. Of these, image registration and segmentation are two crucial components during the analysis process. For instance, image registration is often needed to align different imaging modalities of a subject to provide complementary information for either diagnosis or interventional purposes, and image segmentation is usually employed to extract meaningful quantitative measurements such as lesion loads or structure changes to aid the follow-up diagnosis. Accurate registration and segmentation are very important for clinical studies. Though these have been researched for decades, improvements in terms of both accuracy and speed are still very necessary.

In this thesis, we focus on investigating machine learning solutions to improve MR image

reconstruction and analysis. In particular, this thesis focuses on three main parts: brain lesion segmentation, dynamic MR image reconstruction and image registration. Our contributions in this thesis can be summarised as follows:

A large margin algorithm for automated segmentation of whiter matter hyperintensity (WMH) in brain MRI. Precise detection and quantification of WMHs is of great interest in studies of neurological and vascular disorders. We propose a novel method for automatic WMH segmentation with both supervised and semi-supervised large margin algorithms provided by the framework. The proposed algorithms optimise a kernel based max-margin objective function which aims to maximise the margin between inliers and outliers. It is also shown that the semi-supervised learning problem can be formulated to learn a classifier and label assignment simultaneously, which can be solved efficiently by an iterative algorithm. Specifically, the proposed model is learnt first via the supervised approach and then fine-tuned on a target image by using the semi-supervised algorithm. We evaluate our method on 88 brain FLAIR MR images from subjects with vascular disease. Quantitative evaluation of the proposed approach shows that it outperforms other well known methods for WMH segmentation.

A CNN-based framework for the differential segmentation of WMHs and stroke lesions. The manual delineation of WMHs is a very tedious, costly and time consuming process, which needs to be carried out by expert annotators. The problem of WMH delineation is further complicated by the fact that other pathological features (i.e. stroke lesions) often also appear as hyperintense regions. However, a task specific, reliable, fully automated method that can segment and differentiate between these two pathological manifestations on MRI has not yet been fully identified. Here we propose to use a fully convolutional CNN architecture, called uResNet, to segment hyperintensities and differentiate between WMHs and stroke lesions, while considering the class imbalance issue. Specifically, we aim to distinguish between WMH pathologies from those caused by stroke lesions due to either cortical, large or small subcortical infarcts. The proposed CNN architecture is shown to outperform other well established and state-of-the-art algorithms in terms of overlap with manual expert annotations.

A convolutional recurrent neural network model for the dynamic cardiac MR

image reconstruction. The key ingredient to the dynamic MR image reconstruction problem is to exploit the temporal correlations of MR sequences for the removal of aliasing artefacts. Here we propose a unique, novel convolutional recurrent neural network (CRNN) architecture which reconstructs high quality cardiac MR images from highly undersampled k -space data by jointly exploiting the dependencies of the temporal sequences as well as the iterative nature of the traditional optimisation algorithms. In particular, the proposed architecture embeds the structure of the traditional iterative algorithms in a learning setting, efficiently modelling the recurrence of the iterative reconstruction stages by using recurrent hidden connections over such iterations. In addition, spatio-temporal dependencies are simultaneously learnt by exploiting bidirectional recurrent hidden connections across time sequences. The proposed method is able to learn both the temporal dependency and the iterative reconstruction process effectively with only a very small number of parameters, while outperforming current MR reconstruction methods in terms of reconstruction accuracy and speed.

A CNN-based network termed k - t NEXT for the dynamic MR image reconstruction via exploiting spatio-temporal correlations. Dynamic MRI exhibits high correlations in k -space and time. In order to accelerate the dynamic MR imaging and to exploit k - t correlations from highly undersampled data, here we propose a novel deep learning based approach for dynamic MR image reconstruction, termed k - t NEXT (k - t NEtwork with X- f Transform). In particular, inspired by traditional methods such as k - t BLAST and k - t FOCUSS, we propose to reconstruct the true signals from aliased signals in x - f domain to exploit the spatio-temporal redundancies. Building on that, the proposed method then learns to recover the signals by alternating the reconstruction process between the x - f space and image space in an iterative fashion. This enables the network to effectively capture useful information and jointly exploit spatio-temporal correlations from both complementary domains.

A joint learning framework for cardiac motion estimation and segmentation. Cardiac motion estimation and segmentation play important roles in quantitatively assessing cardiac function and diagnosing cardiovascular diseases. We propose a novel deep learning method for joint estimation of motion and segmentation from cardiac MR image sequences. The proposed network consists of two branches: a cardiac motion estimation branch which is

built on a novel unsupervised Siamese style spatial transformer network, and a cardiac segmentation branch that is based on a fully convolutional network. The joint multi-scale feature encoder is learnt by optimising the segmentation branch and the motion estimation branch simultaneously. The framework is also extended to joint predictions directly from undersampled MR images, which enables to extract clinically relevant measures while bypassing the usual image reconstruction stage.

An unsupervised deep learning based deformable image registration for multi-modal images. Multi-modal registration is a key problem in many medical image analysis applications. It is very challenging due to complicated and unknown relationships between different modalities. We propose an unsupervised learning approach to reduce the multi-modal registration problem to a mono-modal one through image disentangling. In particular, images of both modalities are disentangled into a common latent shape space and separate latent appearance spaces, and the proposed registration approach is then built on the factorised latent shape code, with the assumption that the intrinsic shape deformation is preserved in this latent space. Specifically, two metrics have been proposed for training the proposed network: a latent similarity metric defined in the common shape space and a learning-based image similarity metric based on an adversarial loss. Results show that our proposed methods achieve competitive performance against other methods at substantially reduced computation time.

1.3 Thesis Overview

The research presented in this thesis mainly focuses on three parts: brain lesion segmentation, dynamic MR image reconstruction and image registration. The remainder of the thesis is organised as follows:

Chapter 2 provides an overview of machine learning methods including both classical machine learning methods and deep learning methods that are relevant to the work in this thesis.

Part I focuses on the brain lesion segmentation, which consists of Chapter 3, Chapter 4 and Chapter 5. Chapter 3 introduces the background of brain MRI white matter hyperintensity

(WMH) segmentation and the related work. Chapter 4 presents a large margin algorithm for the automated segmentation of WMHs in brain MRI, which effectively utilises both labelled and unlabelled information. In subjects with vascular diseases, WMHs often coexist with stroke lesions, which makes the WMH segmentation difficult. Thus in Chapter 5, a method for differential segmentation of WMHs and stroke lesions is presented, which is based on a U-shaped deep learning model with residual blocks.

Part II presents the research on dynamic MR image reconstruction, including Chapter 6, Chapter 7 and Chapter 8. In Chapter 6, an overview of background on MRI reconstruction and its related work are provided. Chapter 7 introduces a novel convolutional recurrent neural network architecture for the dynamic cardiac MRI reconstruction, exploiting the temporal redundancies between frames. In order to further exploit the spatio-temporal correlations, a k - t network with x - f transformation is proposed in Chapter 8, which recovers the images in both x - f and image domains.

Part III mainly focuses on the image registration and motion estimation. Similarly, Chapter 9 provides an overview of the deformable image registration techniques, including both mono-modal and multi-modal image registration. Chapter 10 introduces a joint framework for motion estimation and segmentation of cardiac MRI, on both fully-sampled and undersampled images, where it shows that these two tasks are complementary. In Chapter 11, the work on cardiac motion estimation presented in Chapter 10 is further extended to an unsupervised multi-modal deformable image registration, which is achieved via embedding disentangled representations onto a common domain.

Finally, Chapter 12 concludes the work presented in this thesis and discusses some limitations and potential future work.

1.4 Publications

The work of this thesis is mainly based on the following research publications:

1. **C. Qin**, R. Guerrero, C. Bowles, C. Ledig, P. Scheltens, F. Barkhof, H. Rhodius-Meester, B. Tijms, A. W. Lemstra, W. M. van der Flier, B. Glocker, D. Rueckert. **A Semi-supervised Large Margin Algorithm for White Matter Hyperintensity Segmentation.** In International Workshop on Machine Learning in Medical Imaging. Springer, 2016: 104-112.
2. **C. Qin**, R. Guerrero, C. Bowles, L. Chen, D. A. Dickie, M.C. Valdés-Hernández, J. Wardlaw, D. Rueckert. **A Large Margin Algorithm for Automated Segmentation of White Matter Hyperintensity.** Pattern Recognition, 2018, 77: 150-159.
3. **C. Qin**, R. Guerrero, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M.C. Valdés-Hernández, D.A. Dickie, J. Wardlaw, D. Rueckert. **White Matter Hyperintensity and Stroke Lesion Segmentation and Differentiation Using Convolutional Neural Networks.** NeuroImage Clinical, 2018, 17: 918-934.
4. **C. Qin**, J. Schlemper, J. Caballero, J. Hajnal, A. Price and D. Rueckert. **Convolutional Recurrent Neural Networks for Dynamic MR Image Reconstruction.** IEEE Transactions on Medical Imaging, 38(1), 280-290, 2019.
5. **C. Qin**, J. Schlemper, J. Duan, A. Price, J. Hajnal and D. Rueckert. **k -t NEXT: Dynamic MR Image Reconstruction Exploiting Spatio-temporal Correlations.** In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2019: 505-513.
6. **C. Qin**, W. Bai, J. Schlemper, S. Petersen, S. Piechnik, S. Neubauer and D. Rueckert. **Joint Learning of Motion Estimation and Segmentation for Cardiac MR Image Sequences.** In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018: 472-480.
7. **C. Qin**, W. Bai, J. Schlemper, S. Petersen, S. Piechnik, S. Neubauer and D. Rueckert. **Joint Motion Estimation and Segmentation from Undersampled Cardiac MR Image.** International Workshop on Machine Learning for Medical Image Reconstruction, 2018: 55-63.

8. **C. Qin**, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen. **Unsupervised Deformable Registration for Multi-Modal Images via Disentangled Representations**. In International Conference on Information Processing in Medical Imaging (IPMI), 2019: 249-261.

Here is an extended list of publications based on collaborations. These works also contribute to MR image reconstruction and analysis, though they are not included in this thesis.

1. C. Bowles, **C. Qin**, C. Ledig, R. Guerrero, R. Gunn, A. Hammers, E. Sakka, D. A. Dickie, M.C. Valdés-Hernández, N. Royle, J. Wardlaw, H. Rhodius-Meester, B. Tijms, A. W. Lemstra, W. M. van der Flier, F. Barkhof, P. Scheltens, D. Rueckert. **Pseudo-healthy Image Synthesis for White Matter Lesion Segmentation**. In International Workshop on Simulation and Synthesis in Medical Imaging. Springer, 2016: 87-96.
2. C. Bowles, **C. Qin**, R. Guerrero, R. Gunn, A. Hammers, D. A. Dickie, M.C. Valdés-Hernández, J. Wardlaw, D. Rueckert. **Brain Lesion Segmentation through Image Synthesis and Outlier Detection**. NeuroImage Clinical, 2017, 16(C).
3. W. Bai, H. Suzuki, **C. Qin**, G. Tarroni, O. Oktay, P. M Matthews, D. Rueckert. **Recurrent Neural Networks for Aortic Image Sequence Segmentation with Sparse Annotations**. In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018: 586-594.
4. J. Schlemper, O. Oktay, W. Bai, D. Castro, J. Duan, **C. Qin**, J. Hajnal and D. Rueckert. **Cardiac MR Segmentation from Undersampled k -space using Deep Latent Representation Learning**. In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018: 259-267.
5. J. Schlemper, D. C Castro, W. Bai, **C. Qin**, O. Oktay, J. Duan, A. N Price, J. Hajnal, D. Rueckert. **Bayesian Deep Learning for Accelerated MR Image Reconstruction**. In International Workshop on Machine Learning for Medical Image Reconstruction, 2018: 64-71.

6. J. Schlemper, J. Duan, C. Ouyang, **C. Qin**, J. Caballero, J. Hajnal, D. Rueckert. **Data Consistency Networks for (Calibration-less) Accelerated Parallel MR Image Reconstruction.** In International Society for Magnetic Resonance in Medicine 27th Annual Meeting & Exhibition, 2019.
7. G. Seegoolam, J. Schlemper, **C. Qin**, A. N Price, J. Hajnal, D. Rueckert. **Exploiting Motion for Deep Learning Reconstruction of Extremely-undersampled Dynamic MRI.** In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2019: 704-712.
8. J. Duan, J. Schlemper, **C. Qin**, C. Ouyang, W. Bai, C. Biffi, G. Bello, B. Statton, D. O'Regan, D. Rueckert. **VS-Net: Variable Spitting Network for Accelerated Parallel MRI Reconstruction.** In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2019: 713-722.

Chapter 2

Background

2.1 Introduction

Machine learning is a specific subset of artificial intelligence (AI) techniques that enable systems to learn and improve from experience without being explicitly programmed. Machine learning algorithms learn to make predictions or decisions based on the observations or data, known as “training data”. They have been extensively used in various applications, such as computer vision and natural language processing.

According to whether labels for the training data are available, machine learning algorithms can be divided into supervised learning and unsupervised learning approaches. In supervised learning settings, the learning algorithms learn the patterns and build computational models based on a set of training examples with inputs and their ground truth labels, such as in classification or regression tasks. By comparing the errors between their outputs and ground truth labels in the training data, such algorithms are able to learn an inference function to make predictions given any new input. In contrast, unsupervised learning methods do not rely on any training labels, and they draw the inferences by exploring the underlying patterns in the unlabelled data set. One of the main applications of unsupervised learning is clustering, where similar observations are grouped into the same cluster. Semi-supervised learning algorithms fall between the supervised learning and unsupervised learning methods, where both labelled

information and unlabelled information are utilised during the training process. They are often employed when labelled data is difficult to acquire and when there is only a limited amount of labelled samples available. In addition, reinforcement learning is a kind of learning methods that allows agents to interact with their environment by taking actions to maximise some cumulative rewards. It allows agents to automatically decide the ideal action within a specific context with the aim to maximise its performance.

In recent years, deep learning methods have been well studied and have become the main stream approaches for computer vision tasks. Deep learning is a subset of machine learning, and it is a specific approach used for building and training neural networks. In contrast to classical machine learning approaches which require the hand-crafted identification of features in data, deep learning methods can discover useful features or patterns in large datasets without human supervision, mimicking a system of human neurons. It has been shown a great success in many computer vision and medical image analysis tasks.

This chapter provides an overview of both some classical machine learning methods and more recent deep learning models, which are relevant to the image analysis presented in this thesis. Some representative traditional machine learning models are described in Section 2.2, and some more recent and popular deep learning approaches are presented in Section 2.3.

2.2 Classical Machine Learning Methods

This section presents some of the representative classical machine learning methods that have been widely used in medical image analysis, including naive Bayes classifier, support vector machine, random forests and artificial neural networks. For more detailed descriptions and explanations, please refer to [34].

2.2.1 Naive Bayes Classifier

A Naive Bayes classifier [172, 279] is a probabilistic machine learning model based on the Bayes' theorem. It is built on the 'naive' assumption that given the value of the class variable, every pair of features are conditional independent. Bayes' theorem is given as:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y)P(x_1, x_2, \dots, x_n|y)}{P(x_1, x_2, \dots, x_n)}, \quad (2.1)$$

corresponding to

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

With the naive conditional independence assumption, Eq. 2.1 can be simplified as

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i|y)}{P(x_1, x_2, \dots, x_n)}. \quad (2.2)$$

Given the input data, $P(x_1, x_2, \dots, x_n)$ remains constant, and thus

$$P(y|x_1, x_2, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i|y). \quad (2.3)$$

Thereby the class y can be determined by

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i|y) \quad (2.4)$$

where the Maximum A Posterior (MAP) estimation can be used to estimate $P(y)$ and $P(x_i|y)$.

There are many types of naive Bayes classifier and they mainly differ in the assumptions on the distribution of $P(x_i|y)$, such as Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes etc. Though naive Bayes classifiers have over-simplified assumptions, they have shown great success in many real-world problems, especially in spam filtering and document classification. And they are extremely fast compared to more sophisticated methods.

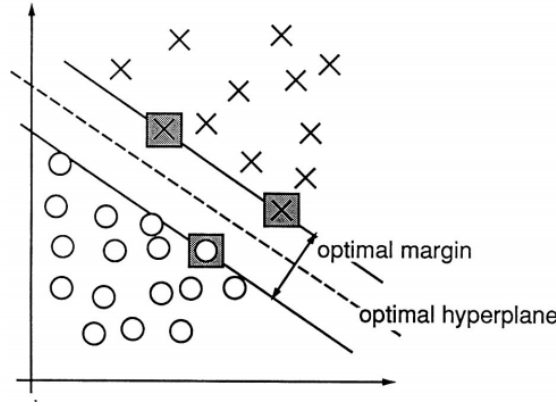


Figure 2.1: Maximum-margin hyperplane and margins for an SVM trained with samples from two classes in 2D feature space. Image adopted from [56].

2.2.2 Support Vector Machines

Support vector machines (SVM) [56] are a popular approach for classification and regression. It is a supervised discriminative method that learns a hyperplane or a set of hyperplanes to separate training data points from different classes by a large margin. An illustrative diagram of SVM in 2D space is shown in Fig. 2.1. As indicated in Fig. 2.1, if the training data is linearly separable, two hyperplanes can be found to separate the two classes of data, and the distance between the two hyperplanes are called margin. The objective of SVM is to find a hyperplane that can best represent the largest margin between the two classes, and the maximum-margin hyperplane is the one that lies halfway between them. To note, samples on the margin are called the support vectors.

Mathematically, given training data $(\mathbf{x}_i, y_i)$ for $i = 1, \dots, n$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$, in the binary classification case, SVM aims to learn a linear classifier $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ to separate the two classes such that $f(\mathbf{x}_i) \geq 1$ if $y_i = 1$, and $f(\mathbf{x}_i) \leq -1$ if $y_i = -1$, i.e., $y_i f(\mathbf{x}_i) \geq 1$. Therefore, the objective function of SVM classifier can be formulated as:

$$\min_{\mathbf{w}} \|\mathbf{w}\|^2, \quad \text{s.t.}, \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \quad \text{for } i = 1, \dots, n. \quad (2.5)$$

However, in cases where data points are not linearly separable, an extension of SVM is to consider the trade-off between the margin and the number of misclassified training data. Then

a slack variable ξ is introduced which allows the margin violation, and the objective function considering the soft margin then becomes:

$$\min_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 + \frac{1}{n} \sum_{i=1}^n \xi_i, \quad s.t., \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \quad \text{for } i = 1, \dots, n. \quad (2.6)$$

Here the primal problem can be solved by converting it to its dual problem, and its Lagrangian dual is:

$$\begin{aligned} \max_{\alpha \geq 0} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i \alpha_i (\mathbf{x}_i, \mathbf{x}_j) y_j \alpha_j \\ s.t., \quad & \sum_{i=1}^n \alpha_i y_i = 0, \text{ and } 0 \leq \alpha_i \leq \frac{1}{2n\lambda}, \quad \text{for } i = 1, \dots, n. \end{aligned} \quad (2.7)$$

This dual problem can be solved via quadratic programming algorithms, and the dual version of classifier then turns to

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i y_i (\mathbf{x}_i^\top \mathbf{x}) + b, \quad (2.8)$$

where b can be recovered by finding an $\mathbf{x}_i$ on the margin's boundary and solving $y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$. To note, α_i is only non-zero on support vectors $\mathbf{x}_i$.

Linear SVM can be extended to learn a non-linear classifier by using a kernel trick [9, 36]. A kernel function is defined as $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$, and the most used kernel functions are Gaussian kernel, RBF kernel, etc. By replacing the linear multiplication with the kernel function, the non-linear classifier can be written as:

$$f(\mathbf{x}) = \sum_{i=1}^N \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b. \quad (2.9)$$

For more details about its derivation, please refer to [36].

2.2.3 Random Forest

Random forest is an ensemble learning method that uses a large number of individual decision trees [39], and the model's prediction is aggregated by combining its decision trees' outputs, such as averaging or majority voting. An illustrative diagram of a random forest is displayed

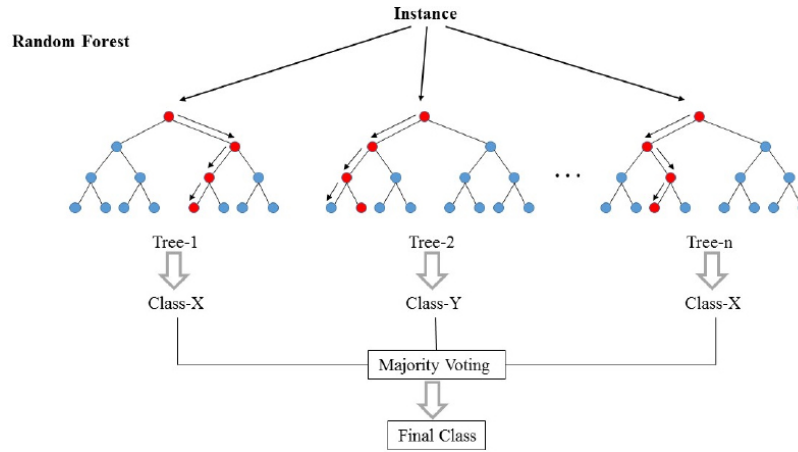


Figure 2.2: An illustrative diagram showing an example of random forest. A random forest composes of a number of decision trees. The final decision of the random forest is made by the majority voting from its decision trees. Image adopted from [69].

in Fig. 2.2. In general, decision trees are a type of model that use a tree-like graph for both classification and regression. It consists of a root node and several internal nodes and leaf nodes, where each internal node corresponds to a ‘test’ on an attribute, and each leaf node represents a decision. The flow of the tree begins from the root node, followed by some sequential questions along a certain route of the tree and finally reaches the prediction. The key to train a decision tree is the strategy to split the attribute space maximising the possibility that samples at the same non-leaf node are from the same class. This split is often achieved based on the computation of information gain.

A random forest consists of multiple decision trees, and this is based on the concept that combination of different uncorrelated models can outperform any of the individual one. There are two types of randomness in the random forest: firstly each decision tree only has access to a random subset of training data, and secondly it only considers a random subset of features to generate the best split. This introduces more diversities in the forest and can produce more robust overall predictions. Random forests have been widely applied in different domains. They do not need feature scaling, categorical feature encoding, and there is little need of parameter tuning. In contrast to some other more complex models such as neural networks, random forests can be more interpretable and accurate.

2.2.4 Artificial Neural Network

Artificial neural network (ANN) is a computational model inspired by the biological neural networks of the brain. It is made up of a collection of connected units called neurons that transform the input to output. The concept of ANN was first proposed in 1940s by McCulloch and Pitts [175], in which the inputs were one or more binary values, and the output was activated when certain inputs were True. Later, Rosenblatt [207] presented a modified version of neurons called ‘Perceptron’, as shown in Fig. 2.3. In contrast to [175], here the inputs x_i are associated with weights w_i , and the summation of input-weight products are fed to an activation function σ , which determines to what extent the signal should pass through the network. The final output then can be expressed as:

$$y_i = \sigma\left(\sum_{i=1}^m w_i x_i + w_0\right), \quad (2.10)$$

where the weights w_i and bias w_0 can be learnt given the training data. By stacking multiple Perceptrons in rows, it then becomes multiple layer Perceptrons (MLP), as presented in Fig. 2.4. Here the output of each neuron is a linear combination of its inputs, and each layer’s output is its subsequent layer’s input. In Fig. 2.4, it shows a neural network architecture which contains two hidden layers, and one input and output layer.

The backpropagation (BP) algorithm is the key element in training artificial neural networks. It was originally introduced in [270], and was later popularised by Rosenblatt [207] to train an MLP network. The BP algorithm computes the gradients based on the chain rule, and then gradient descent approaches can employ such gradients to update the associated weights and optimise loss functions. The training algorithm can be described mainly in four steps: 1) given the input data, compute the predictions through the feedforward pass; 2) calculate the loss errors; 3) backpropagate the errors from the final layers to previous layers by repeatedly using chain rule; 4) and update weights and biases using gradient descent approaches. It has now become a basic for training MLP and even deeper neural networks.

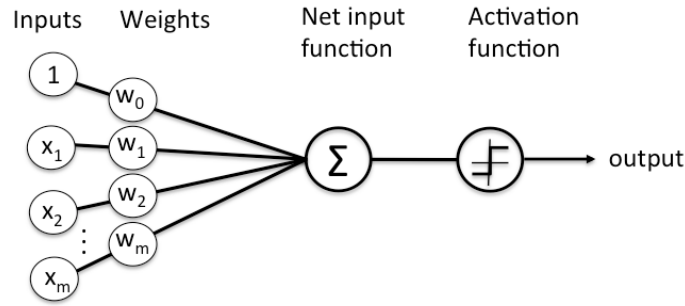


Figure 2.3: An illustrative diagram of an artificial Perceptron. The output of the Perceptron is a weighted sum of the inputs followed by a linear threshold function. Image taken from [229].

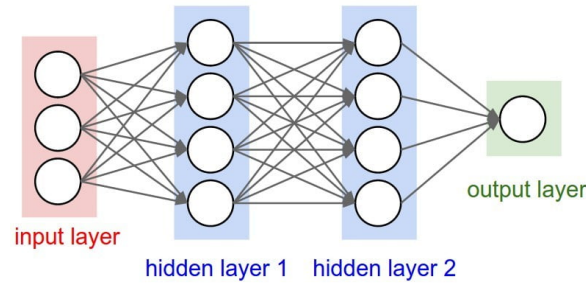


Figure 2.4: An illustrative diagram of a multi-layer Perceptron, with an input layer, two hidden layers and an output layer. Image adopted from <http://cs231n.github.io>.

2.3 Deep Learning Methods

In recent years, research on deep learning methods is developing rapidly. Deep learning is a subset of machine learning methods based on artificial neural networks. It uses multiple and deep layers to model the complex relationships between inputs and outputs. Convolutional neural networks (CNNs), recurrent neural networks (RNNs) and generative adversarial networks (GANs) are some of the most popular deep learning models that have been actively researched and used. Here we will focus on the ones that will be used in the following chapters. There are also many other deep learning models, such as variational autoencoder [70], deep reinforcement learning [83], bayesian deep learning [258], capsule network [210] and many others. For detailed information about more deep learning methods, please refer to [94].

2.3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) has raised a lot of interests in recent years and has been shown to be powerful in computer vision tasks. It is similar to the original artificial neural networks, where each neuron receives inputs, and performs a dot product followed by an optional non-linearity. Different from regular neural networks, CNNs architecture explicitly assumes that inputs are images, and neurons at each layer are only connected to a small region of the previous layer, instead of a fully-connected manner.

The standard architecture of CNNs was inspired by the hierarchical connectivity pattern between neurons of visual systems. A simple CNN architecture usually consists of three main types of layers: convolutional layer, pooling layer and fully connected layer.

- **Convolutional layer:** The convolutional layer is the core element in convolutional neural networks. The layer consists of a set of learnable filters with learnable weights and biases. The filter is often 3D with a small spatial extent defined by the width and height, and expands along the depth dimension of the whole input volume. As mentioned earlier, each neuron in a convolutional layer is only connected to a small region of the input volume, and the spatial extent of such connectivity is a hyperparameter called the receptive field of the neuron. For instance, a typical kernel size of the convolutional filters at the first convolutional layer is $5 \times 5 \times 3$, which means that it spatially convolves 5 pixels of the image along width and height dimensions, and 3 along its channel axis. During the forward pass, each filter slides across the whole input volume along width and height, and then the dot product is computed between the filter weights and the input volume. This produces an activation map which corresponds to responses of filters at each spatial position, and the output volume is generated by stacking the activation maps produced from a set of filters along the channel axis. To note, in contrast to original neural networks whose neurons at each layer are independent, here the weights are shared for each neuron, with the assumption that if one feature is useful to compute at some spatial location, then it should be also useful at other positions.

- **Pooling layer:** The pooling layer is common in a CNN architecture, which reduces the spatial size of the representation and thus reduces the number of parameters needed in subsequent layers. Pooling layers are often inserted in between convolutional layers, and the max-pooling is the most common operation used which outputs the maximum of a small region from each depth independently, such as a 2×2 size with a stride of 2. In addition to the max pooling, average pooling is also often used, but max pooling has gained the popularity over average pooling due to its better performance in practice.
- **Fully connected layer:** The fully connected layer is fully connected to all activations from previous layer, and it is the same with the regular neural networks. The activation is computed by the matrix multiplication with a bias offset. In recent works, fully connected layers are often converted to a convolutional layer which enables the network to be fully convolutional thus makes it not limited to the input image size. For instance, a fully connected layer with $K = 1024$ that looks at an input volume of size $5 \times 5 \times 256$ can be equivalently converted to a convolutional layer with filter size of $5 \times 5 \times 256$ with number of filters of 1024.

In recent years, there are many variants of CNN architectures that have been proposed for image classification or image segmentation tasks. One of the first successful applications of CNN is LeNet, which was developed by LeCun et al. [149] in 1990s. The detailed architecture of the LeNet-5 is shown in Fig. 2.5. It was used for handwritten digits recognition and has been applied commercially. Based on LeNet, AlexNet [143] was developed and popularised CNN for computer vision tasks. AlexNet has a similar architecture with LeNet, but is deeper and bigger, and features convolutional layers stacked on top of each other. The architecture of AlexNet is shown in Fig. 2.6. It won the ImageNet Large Scale Visual Recognition Competition (ILSVRC) in 2012, and outperformed the second runner-up significantly. In ILSVRC 2013 and ILSVRC 2014, ZF Net (Zeiler Fergus Net) [278] and GoogLeNet [237] became the winners of the challenge respectively. An Inception Module was proposed in GoogLeNet that reduced the amount of parameters dramatically. The runner-up of ILSVRC challenge in 2014 was the VGGNet [228], in which it showed that the depth of the network plays an important role in

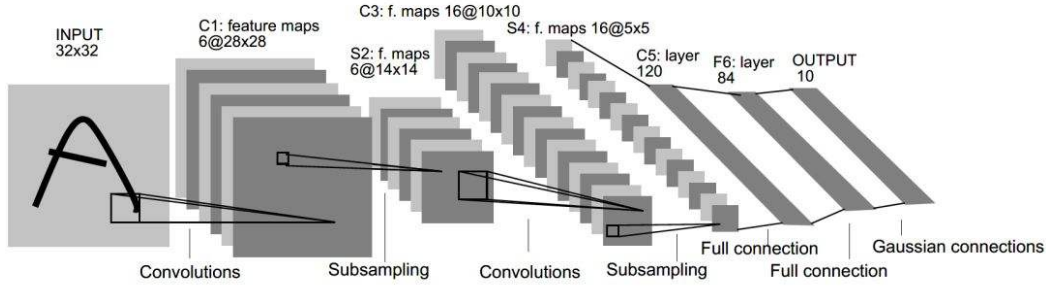


Figure 2.5: The network architecture of LeNet-5 for handwritten digits recognition. Image taken from [149].

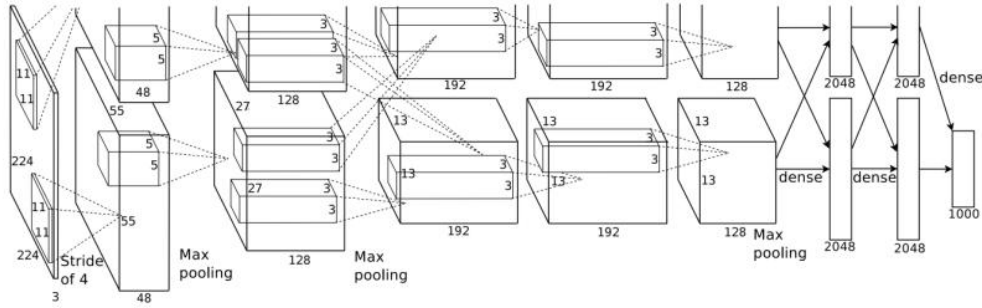


Figure 2.6: The network architecture of AlexNet. Image taken from [143].

achieving good performance. An example of the VGGNet (VGG-16) is shown in Fig. 2.7, where it contains 16 convolutional or fully connected layers, with 3×3 and 2×2 pooling across the whole network.

In more recent years, ResNet [107] has gained its popularity and became the default choice for using convolutional networks. It was proposed by He et al. and was the winner of ILSVRC 2015. In particular, ResNet contains the residual blocks as shown in Fig. 2.8, where it has two paths: one is the regular convolutional path, and the other is a skip connection that corresponds to an identity mapping. Such residual connection addresses the problem of gradient vanishing, and also the problem of performance degradation for deep networks. This residual design enables the network to get deeper while achieving better performance, and the depth can be up to 152 layers, $8\times$ bigger than VGGNet. Later, a DenseNet architecture [116] was further proposed to improve the performance of ResNet, in which layers in a dense block are densely connected, as shown in Fig. 2.9.

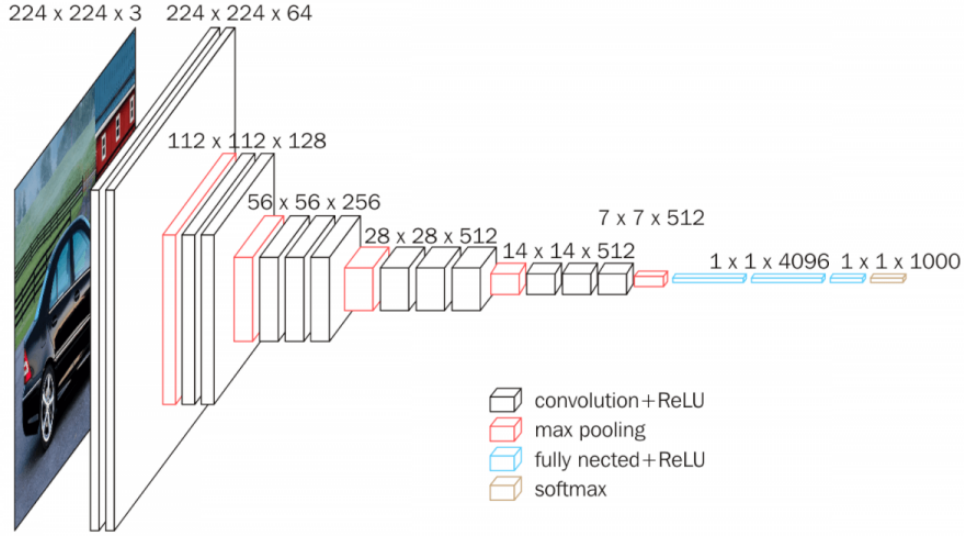


Figure 2.7: The network architecture of VGG-16. Image adopted from [228].

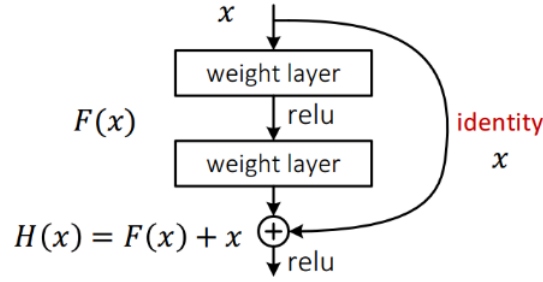


Figure 2.8: The architecture of residual block in ResNet. Image adopted from [107].

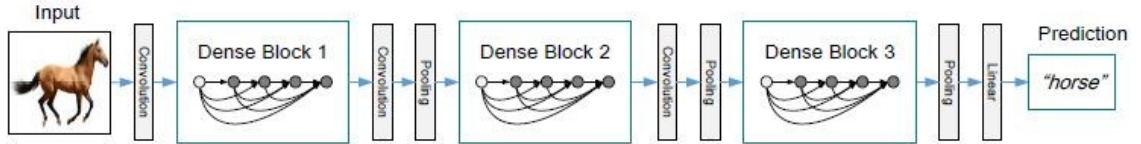


Figure 2.9: The network architecture of DenseNet. Image adopted from [116].

Most of the above mentioned methods are often used to solve the image classification task. In terms of image segmentation, a fully convolutional network (FCN) was proposed by Long et al. [160], in which the semantic segmentation were generated by combining feature maps from different scales, as shown in Fig. 2.10. To note, here in FCN, there is no use of fully connected layers, which enables the network to predict segmentations for images of any size, thus is more efficient than the patch-based approaches. This work inspires most of the follow-

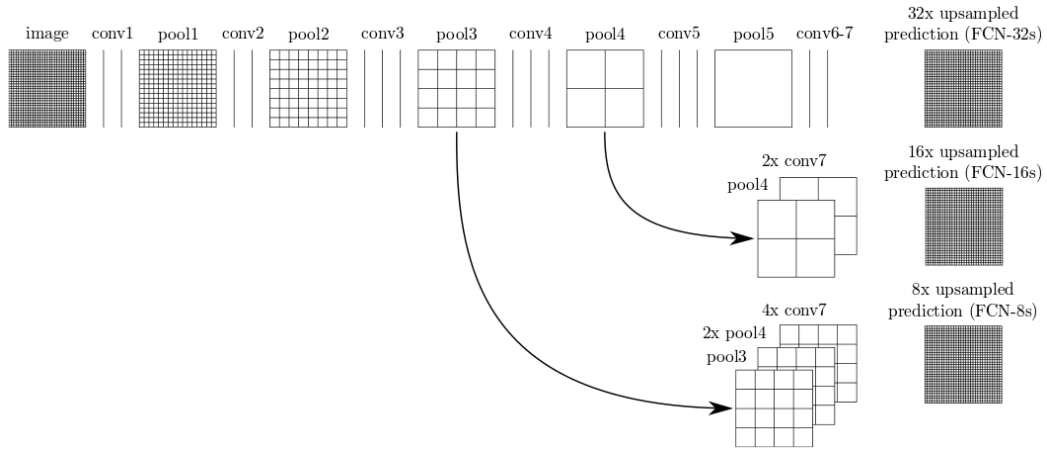


Figure 2.10: The network architecture of FCN for image semantic segmentation. Image taken from [160].

up work on semantic segmentation. Some other popular architectures for segmentation include SegNet [19] and DeconvNet [183], which have an encoder-decoder architecture with symmetric convolutional and deconvolutional layers, and pooling and upsampling/unpooling layers. One of the widely used architecture for medical image segmentation is U-net proposed by Ronneberger et al. [206], in which feature maps from analysis path are concatenated with feature maps from synthesis path for the predication of segmentations. This allows the gradients from higher layers to be propagated to lower layers directly, which has been shown to be beneficial for the performance gain. Variants of U-net have also been proposed for image segmentation, such as 3D U-net [54] and Tiramisu Net [124].

2.3.2 Recurrent Neural Networks

Recurrent neural networks (RNNs) are a class of neural networks that make use of sequential information to process sequences of inputs, where units in a network are connected to form a directed cycle. In recent years, RNNs have also been widely researched and applied in various applications such as natural language processing, image captioning, video analysis etc.

In contrast to normal neural networks, RNNs was proposed to remember the previous information while processing the current data. It maintains an internal state (hidden state) to propagate information between sequential steps. The network is formed as a loop, as shown

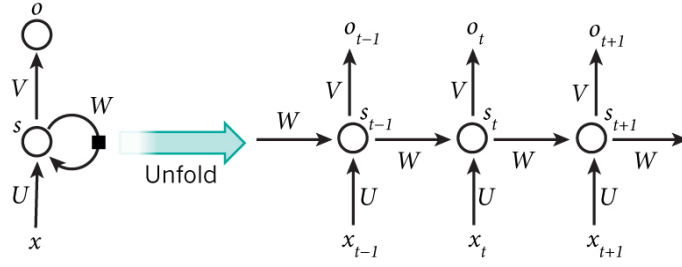


Figure 2.11: Diagram of an unrolled recurrent neural networks. Image adopted from [148].

in Fig. 2.11, which allows the information to be passed from one step to another. Specifically, the network receives input x , and outputs an output o and a hidden state s . If the network is unfolded, it becomes a chain of networks with shared weights at each step, where each unit receives input from a sequence of time points with hidden states propagated along the sequential steps. This can be described in a mathematical form:

$$s_t = \phi(Ux_t + Ws_{t-1}), \quad (2.11)$$

where s_t is the hidden state at time step t , and it is the function of the input at current step x_t multiplied by a weight matrix U , with an addition to a hidden state from the previous time step s_{t-1} weighted by a hidden-to-hidden matrix W . The weight matrices determines how much information from the previous hidden state and the current input should contribute to the current process. This is then followed by a non-linearity function ϕ that squashes the output to a certain range. A diagram of the unfolded RNNs is shown in Fig. 2.11. RNNs enable the network to model the short term memory, with sequential information preserved in hidden states.

However, one problem with the vanilla RNNs is the exploding or vanishing gradient problem. Optimisation of recurrent networks relies on the backpropagation through time, and computation of gradients requires many stages of multiplication through layers and time steps in deep neural networks. This could potentially lead to the derivatives susceptible to exploding or vanishing. Common practice to address the exploding gradient problem is to employ gradient clipping, truncating the gradients to a pre-defined small range, and rectified linear unit (RELU) is often used to avoid the vanishing gradient problem.

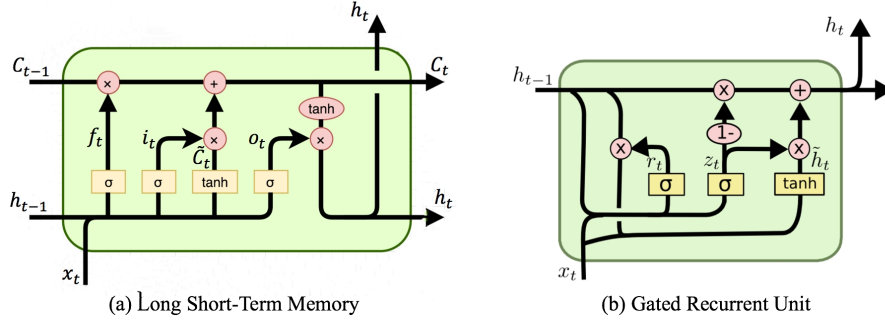


Figure 2.12: Illustration diagram of LSTM and GRU unit. Images taken from <https://colah.github.io/>.

Long Short-Term Memory Units

Long Short-Term Memory Units (LSTM) is a special kind of RNNs that is capable of learning long-term dependencies. It was first proposed in mid-90s by Hochreiter and Schmidhuber [111] as a solution to the vanishing gradient problem, which enables information to be memorised for long periods of time. The structure of the unit is shown in Fig. 2.12.

The core idea of LSTM is the cell state C , which contains information outside the normal flow of the recurrent network. LSTM is able to add or remove information from a cell state, which is controlled by structures called gates. Gates act on signals they receive and decide how much information can go through. They are composed of a sigmoid layer with an element-wise multiplication, as shown in Fig. 2.12. The mathematical form of LSTM can be described as follows:

$$f_t = \sigma(U_f x_t + W_f h_{t-1} + b_f) \quad (2.12a)$$

$$i_t = \sigma(U_i x_t + W_i h_{t-1} + b_i) \quad (2.12b)$$

$$\tilde{C}_t = \tanh(U_C x_t + W_C h_{t-1} + b_c) \quad (2.12c)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t \quad (2.12d)$$

$$o_t = \sigma(U_o x_t + W_o h_{t-1} + b_o) \quad (2.12e)$$

$$h_t = o_t \circ \tanh(C_t) \quad (2.12f)$$

In details, the first component of LSTM is the forget gate, which decides how much information

to forget from the previous cell state C_{t-1} by looking at current input x_t and previous hidden state h_{t-1} . The mathematical form of the forget gate is presented in Eq. 2.12a. Then it decides what information is going to add into the cell state. This consists of two parts: an input layer (Eq. 2.12b) that decides what information to update, and a tanh layer (Eq. 2.12c) that produces a new candidate value $\tilde{C}_t$ to add to the cell state. These two are then combined to update the cell state, as shown in Eq. 2.12d, by forgetting the things in old state C_{t-1} , and adding the new candidate values $i_t \circ \tilde{C}_t$. Finally, an output gate o_t decides what parts of cell state to output. Here the output hidden state h_t is obtained by multiplying the o_t with the cell state C_t through tanh.

Gated Recurrent Units

There are many variants on LSTM, and one popular variant is the Gated Recurrent Units (GRU) [53]. The model structure is much simpler than LSTM, as shown in Fig. 2.12. The main difference is that GRU combines the forget gate and input gate as a single update gate, and merges the hidden state and cell state. A detailed mathematical explanation of GRU is described as the following:

$$z_t = \sigma(U_z x_t + W_z h_{t-1} + b_z) \quad (2.13a)$$

$$r_t = \sigma(U_r x_t + W_r h_{t-1} + b_r) \quad (2.13b)$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tanh(U_h x_t + W_h (r_t \circ h_{t-1}) + b_h) \quad (2.13c)$$

GRU is getting increasingly popular in recent years, due to its simplicity over LSTM while maintaining the performance.

RNNs have been widely used in computer vision tasks such as object recognition, image captioning and video analysis, due to their success in processing sequential information. Convolutional LSTM was proposed in [273], where it employed convolution operation in all input-to-state and state-to-state transitions, which makes it efficient to use RNNs on images. Convolutional RNNs have also been applied for object recognition [154], saliency detection

[144], super-resolution [118], etc. In addition, Gregor et al. [96] proposed to use LSTM for the image generation task, where it combines the spatial attention mechanism with a sequential variational auto-encoder framework for the iterative reconstruction of complex images. Image captioning is also an active research field that makes use of RNNs to generate natural language description of images [132, 255, 274]. For instance, Karpathy and Li [132] proposed a model based on the combination of CNNs over image regions, RNNs over sentences and a structured objective aligning two modalities through a multi-modal embedding. Besides, video analysis is a natural application of RNNs due to the temporal dependences between frames. Examples include the video super-resolution [118], future frame prediction [161, 232] and motion estimation [187]. Other applications such as video captioning [72, 253] and visual question answering [14, 85] have also been widely researched exploiting the long-term dependencies of RNNs.

2.3.3 Generative Adversarial Networks

Generative adversarial networks (GAN) were first proposed by Goodfellow et al. [95] in 2014 to estimate generative models via an adversarial process. The framework of GAN consists of two competing neural network models: a generator G to generate new data samples, and a discriminator D to distinguish between real and fake samples. The whole network is trained in an adversarial way, corresponding to a minimax two-player game, where the generator learns to generate synthetic data that can fool the discriminator while the discriminator learns to determine whether a sample is from data distribution or model distribution. The steps are repeated during the training process, and both generator and discriminator can gain improvement in their respective jobs after each iteration. A diagram of the framework of GAN is shown in Fig. 2.13.

Here the generative model G is trained to capture the real data distribution and to maximise the probability of D making a mistake, while the discriminator D estimates the probability of a given sample coming from the real dataset. This can be formulated as a minimax game with

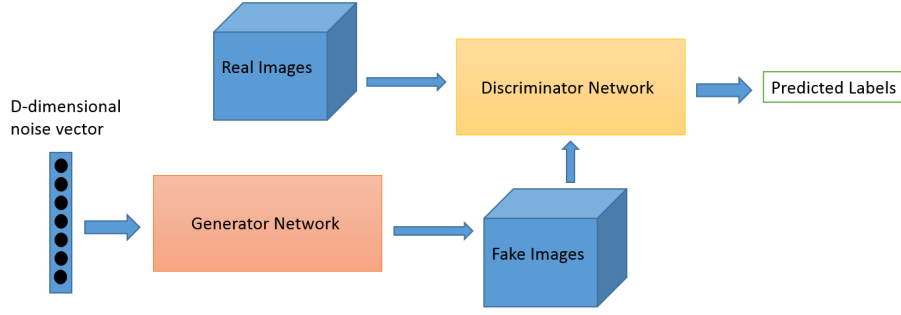


Figure 2.13: The framework of Generative Adversarial Networks. Image adopted from O'Reilly (<https://www.oreilly.com>).

a value function $V(D, G)$:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log(D(x))] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))], \quad (2.14)$$

where p_{data} is the distribution of real data, and p_z is data distribution over noise input z . The training process involves two parts: training of discriminator D while G is idle, and training of generator G while D is idle. For training D , the generator network G is only forward propagated, and the objective is to maximise the probability to discriminate real and fake samples, i.e.,

$$\max_D V(D) = \mathbb{E}_{x \sim p_{data}(x)} [\log(D(x))] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))], \quad (2.15)$$

where the label for real samples is 1, and label for fake samples is 0. On the generator side, its objective is to generate samples with the highest possible value of $D(G(z))$ that can fool the discriminator:

$$\min_G V(G) = \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]. \quad (2.16)$$

The network is then trained in alternating steps until the generator can produce good quality images.

However, there are several limitations with this formulation of GAN. Although GAN has shown great success in generating realistic images, the training process is slow and not stable. Firstly, there is a vanishing gradient problem. If the discriminator is perfect, the loss function

will drop to zero, and there is no gradient flow to update the loss function during learning iterations. It has been shown that when the discriminator gets better, the gradient vanishes fast [15]. In addition, Salimans et al. [212] pointed out that training two models independently with respect to their own cost cannot guarantee a convergence and is hard to achieve the Nash equilibrium. Besides, mode collapse is also a common failure for GAN where generators tend to produce outputs with low variety [15].

Wasserstein GAN

As an alternative to traditional GAN training, Wasserstein GAN (WGAN) was proposed by Arjovsky et al. [15] to improve the learning stability and avoid problems such as mode collapse, where it uses the Wasserstein distance as the new cost function. The Wasserstein distance is a measure of distance between two probability distributions, and its distance formula is:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|]. \quad (2.17)$$

In the formulation, $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ is the set of all joint distributions between the real data distribution $\mathbb{P}_r$ and the generated data distribution $\mathbb{P}_g$.

However, the infimum of the Wasserstein distance is highly intractable. Therefore, [15] proposed to simplify the calculation using the Kantorovich-Rubinstein duality:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r} [f(x)] - \mathbb{E}_{x \sim \mathbb{P}_g} [f(x)]. \quad (2.18)$$

Here f is a 1-Lipschitz function with the constraint:

$$|f(x_1) - f(x_2)| \leq |x_1 - x_2|. \quad (2.19)$$

If there is a family of 1-Lipschitz continuous functions $f_{w \in \mathcal{W}}$ parameterised by w , the discrimi-

nator is then used to learn a function f_w solving the problem:

$$\max_{w \in \mathcal{W}} \mathbb{E}_{x \sim \mathbb{P}_r} [f_w(x)] - \mathbb{E}_{z \sim p_z} [f_w(G(z))]. \quad (2.20)$$

Here the discriminator is not a direct critic to distinguish real and fake samples, and instead it learns a function to help compute the Wasserstein distance. To enforce the 1-Lipschitz continuity constraint, WGAN applies a simple weight clipping to restrict the maximum of weights to a small range. However, the model performance is very sensitive to this clipping hyperparameter, and it still suffers from slow convergence or vanishing gradient problem if the clipping window is too large or too small.

WGAN with gradient penalty

WGAN with gradient penalty (WGAN-GP) [99] has been proposed to improve over the weight clipping for enforcing the Lipschitz constraint. It is based on the theory that a differential function f is 1-Lipschitz if and only if it has gradients with norm at most 1 everywhere. So instead of weight clipping, a gradient penalty term is introduced, and the new objective function then becomes:

$$\mathcal{L} = \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g} [D(\tilde{x})] - \mathbb{E}_{x \sim \mathbb{P}_r} [D(x)] + \lambda \cdot \mathbb{E}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]. \quad (2.21)$$

In this formulation, the last term is the gradient penalty, and to circumvent tractability issues, here the soft version of constraint is enforced along the straight lines between pairs of sampled points from data distribution $\mathbb{P}_r$ and generator distribution $\mathbb{P}_g$, i.e., $\hat{x} = \epsilon x + (1 - \epsilon)\tilde{x}$ with $\epsilon \in U[0, 1]$. Though gradient penalty introduces computational complexity, it shows success in producing higher quality images.

GAN is a very active topic of research and there are many variants and implementations such as Conditional GAN (CGAN) [179], Deep Convolutional GAN (DCGAN) [198], InfoGAN [51], CycleGAN [282] and so on, and they have also been applied in various applications including image generation, image super-resolution, image-to-image translation, etc. DCGAN [198] is one

of the most popular and successful implementation of GAN, where it uses CNNs to replace the multi-layer perceptrons in the vanilla GAN. Conditional GAN [179] adds additional information to generator and discriminator such as class labels to generate corresponding data. Besides, in order to enhance the details of super-resolved images, a Super Resolution GAN (SRGAN) [151] was proposed with a discriminator to differentiate between the super-resolved images and original photo-realistic images. More recently, CycleGAN [282] has become very popular for the unpaired image-to-image translation task, in which no paired images are required during the training process, and the network is trained via a cycle-consistency loss. Similar to that, some works have also investigated on the unsupervised image-to-image translation, such as via learning a joint distribution of images in different domains [158], or through manipulating the disentangled image representations [117, 153]. In the domain of medical image analysis, GANs have also been widely used in various applications [135], such as image reconstruction [220, 280], image registration [80, 165, 238], and image synthesis [57, 182]. For detailed introduction and descriptions of GANs in medical image analysis, please refer to the work [135].

2.3.4 Spatial Transformer Network

Spatial transformer network (STN) was first proposed by [123] to address the limitation of CNNs that have a lack of spatial invariance to large transformations of input data. The proposed differentiable spatial transformer module can be plugged in any neural networks to provide spatial transformation capabilities. The architecture of the spatial transformer module is presented in Fig. 2.14. Specifically, the transformer module is a combination of the localisation network, grid generator and sampler. The localisation network receives input from the feature map U , and outputs the parameter θ for transformation T_θ , i.e., $\theta = f_{loc}(U)$. The size of the parameter θ depends on the form of transformation type, e.g. for a 2D affine transformation θ is 6-dimensional. The architecture of the localisation network f_{loc} can be any form. To perform the warp of the feature map, the regular grid G over the output feature map V is transformed to the sampling grid $T_\theta(G)$ by using the estimated transformation parameter θ . This is then followed by a sampler to produce the sampled output feature map V by warping the input feature map

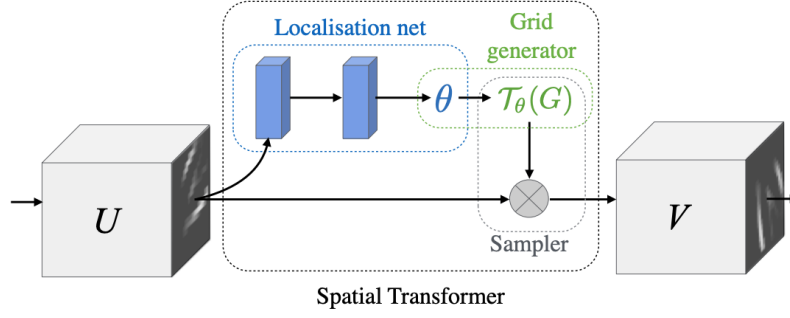


Figure 2.14: The architecture of a spatial transformer module. The input feature map U is passed to a localisation network to regress the transformation parameter θ . The regular grid G over V is transformed to the sampling grid $T_\theta(G)$, which is applied to U to generate the warped feature map V . Image taken from [123].

U . Here any sampling kernel can be used for the sampler, such as integer sampling kernel or bilinear sampling kernel. To note, the sampling is performed identically for each channel of the input, i.e., each channel is transformed in the same way.

The spatial transformer module can be inserted at any place within a CNN network, enabling it to learn how to transform the feature map in order to minimise the overall loss function during the training. Apart from estimating affine transformation and thin plate spline transformation parameters proposed in the original work [123], extensions on dense transformations, i.e., pixel-wise displacements, have also been investigated [187]. In recent years, such framework has been extensively used for unsupervised optical flow estimation [120, 187, 203], in which they trained a STN to estimate the optical flow via minimising the intensity dissimilarities. In the field of medical image analysis, STN was employed for the image registration task to estimate deformation fields between source image and target image, which were learnt via optimising some pre-defined similarity measures [58, 65, 114, 142].

2.4 Conclusions

This chapter has introduced some representative machine learning algorithms, both classical machine learning methods and modern deep learning methods, applied as part of the work

presented in this thesis. Brief introductions of their applications in computer vision and medical image analysis tasks have also been provided. Later chapters will give a more focused reviews of the most recent research on magnetic resonance image (MRI) reconstruction and analysis, including brain lesion segmentation, dynamic MRI reconstruction, and medical image registration.

Part I

Brain Lesion Segmentation

Chapter 3

Background: White Matter

Hyperintensity (WMH) Segmentation

3.1 Introduction

White matter hyperintensities (WMH), referred to in the clinical literature as leukoaraiosis, white matter lesions or white matter disease [266], are a characteristic of small vessel disease [265], which, as the name suggests, appear as hyperintense regions on T2-weighted fluid-attenuated inversion recovery (FLAIR) magnetic resonance (MR) images due to localised, pathological changes in tissue composition [266]. WMHs are commonly observed in elderly subjects and patients with neurodegenerative diseases (NDs), such as vascular dementia (VaD) and Alzheimer’s disease (AD), which are known to be hard to diagnose and have become a major challenge in our society. Damaged white matter usually has a prolonged T2 relaxation time due to increased tissue water content and to degradation of the macromolecular structure of myelin [45]. Therefore, WMHs can be conspicuously depicted with conventional T2-weighted spin echo or fast spin echo sequences, especially on T2-weighted FLAIR images. On the other hand, WMHs tend to be fairly dark on T1 images, as shown in Fig. 3.1, which provides examples of T1 and FLAIR images of one subject showing varying visual characteristic of lesions in different modalities.

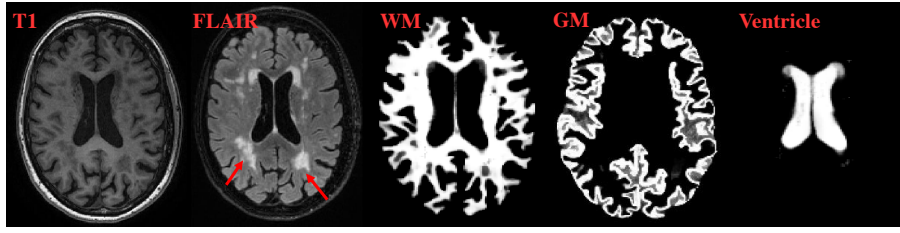


Figure 3.1: An example of T1 image, FLAIR image, WM tissue, grey matter (GM) tissue and ventricle. These images are from corresponding slice of the same subject. Tissues were segmented by an automated brain segmentation tool called MALP-EM [150]. The red arrows point to WMH on the FLAIR.

WMHs in older population may be small, diffuse, and irregular in shape, and sufficiently heterogeneous within and across subjects. Current research [24, 45] indicates that the WMH volume in subjects with dementia is significantly higher than that of a normal aging population, and the presence, severity and distribution of WMHs may also vary between different disorders. Thus, accurate quantification of WMHs in terms of total volume and distribution is believed to be of clinical importance for prognosis, tracking of disease progression and assessment of the treatment effectiveness [88, 192]. WMH volume has been demonstrated to correlate with severity of symptoms, progression of disability and clinical outcome [31, 49, 162]. Accordingly, determining WMH volume has been of interest in clinical research as well as in clinical trials on disease-modifying drugs [3, 42, 162, 250].

Clinically, one of the most widely used metrics to assess WMH burden and severity is the Fazekas visual rating scale (i.e. score) [81]. In this scale, a radiologist visually rates deep white matter and peri-ventricular areas of a MR scan into four possible categories depending on the size, location and confluence of lesions. The combination of both deep white matter and peri-ventricular ratings yields a combined zero to six scale. In the vast majority of clinical trials and in general clinical practice, visual rating scores are used (such as the Fazekas score). However, WMHs are very variable in size, appearance and location, and therefore the categorical nature of the Fazekas scale has limitations for studying their progression in relation with other clinical parameters, and such visual rating scales lack sensitivity for the finer details of subtle differences in WMHs [110].

Manual delineation of WMHs is an alternative way to assess WM abnormalities. For some studies, lesions have been traced manually (sometimes with the help of semi-automated tools for contour detection) slice by slice. However, manual segmentation is a laborious, observer-dependent and time consuming task that is unfeasible for larger datasets and/or clinical practice. Moreover, WMHs in patients with NDs such as VaD can be small, irregular and scattered, which makes the precise segmentation of WMHs rather difficult to tackle. Thus, steps towards more reliable and precise WMH identification and quantification are highly desirable.

On the other hand, stroke lesions of cortical, large subcortical (striatocapsular) or small subcortical infarct origin can often coexist and coalesce with WMHs on patients with stroke or dementia, which also appear as hyperintense regions in FLAIR MR images. Such stroke lesions are similar to WMHs, and could be misclassified as WMHs in some image processing algorithms. However, in the assessment of WMH burden, it is very important to exclude stroke lesions as they have different underlying pathologies. Failure to account for this may lead to an inaccurate calculation of WMH volume, cerebral atrophy and their longitudinal progression measurements [261]. This will thus have effect on the observational studies or randomised trials which use WMH volume and progression as measures. Therefore, methods that can effectively discriminate between WMHs and stroke lesions are very necessary.

Recently, several automated and semi-automated methods have been put forward to address the coarseness of the visual assessments (e.g. Fazekas score), as well as the dependence on highly qualified experts to perform such assessments. These methods can be broadly classified into *supervised*, when a training “gold-standard” is available [88, 252], i.e. when one or more human experts have annotated data, *unsupervised*, when no such gold-standard exists [38, 48, 277], and *semi-supervised*, when only a small portion of available data has been expertly annotated [134, 195]. However, despite the number of proposed methods, no automated solution is currently widely used in clinical practice and only a few of them are publicly available [60, 218, 226]. In addition, in most previously proposed automatic WMH segmentation algorithms, they do not take into account the presence of stroke lesions, and these methods are generally unable to differentiate between these two types of lesions.

3.2 General preprocessing steps

Prior to segmentation, a few image preprocessing steps are generally taken in order to minimise the effect of imaging artefacts and align the different sequences in the same space. The main steps applied prior to the segmentation procedure are the following:

- **Registration:** The goal of registration is to estimate spatial correspondences between images and transform them into one coordinate system. For methods that use more than one sequence, multiple modalities such as T1, T2, and FLAIR need to be registered into a common space before analysis. Registration can also be employed to align an atlas to the brain to provide some initial estimates of the brain tissues.
- **Brain extraction:** Brain extraction is to remove the skull from a brain image, leaving only the regions with actual brain tissues. Segmentation then can be performed only on the remaining brain voxels.
- **Tissue segmentation:** Tissue segmentation is to classify brain tissues into different tissue classes using high-resolution T1-weighted anatomical images. It is often to separate them into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF), but more classes are possible. WMH segmentation is commonly implemented on WM.
- **Bias correction:** Intensity inhomogeneity which are caused by magnetic settings, patients' position, and other factors in MR images are quite common. Bias correction can help to reduce the smooth variation of intensity of the tissues to simplify subsequent segmentation.
- **Intensity normalisation:** Intensity normalisation is a process to change the range of image intensity values. Some segmentation methods require the intensity of the test images to be similar to that of the training images and thus intensity normalisation step is needed.

3.3 Related Work

Over the past decades, there have been an increased number of works in the area of brain MRI analysis [45, 86, 166, 202], including brain and tissue segmentation [126, 204, 208] and abnormality detection in brain MRI [199, 276]. Recently, several techniques that seek to automatically and precisely segment and quantify WMHs have been put forward [45]. In the following, we review existing methods and challenges that are related to our work, especially on Multiple sclerosis (MS), WMH and stroke lesion segmentation in MR imaging. Although some of the methods mentioned here were proposed for segmenting different pathologies rather than WMHs and stroke lesions, they can in fact be applied to different tasks. We categorise these methods into two main groups: traditional methods and deep learning based methods.

3.3.1 Traditional Methods

Arguably, most state-of-the-art approaches for brain lesion segmentation are machine learning-based methods. They can be broadly classified into *unsupervised*, *semi-automatic*, *semi-supervised* and *supervised* methods, depending on the amount of available expertly annotated data.

Unsupervised segmentation

Unsupervised segmentation methods do not require labeled data to perform the segmentation. Most of these approaches employ clustering methods based on intensity information or some anatomical knowledge to group similar voxels into clusters, such as fuzzy C-means methods [91], EM-based algorithms [76, 82, 137] and Gaussian mixture models [84, 136]. Some of the current well known methods will be briefly reviewed in the following.

Lesion Growth Algorithm Recently, a lesion growth algorithm (LGA) [218] has been proposed, and has also been translated into end user software called lesion segmentation toolbox (LST) (<http://www.applied-statistics.de/lst.html>). The proposed algorithm was developed

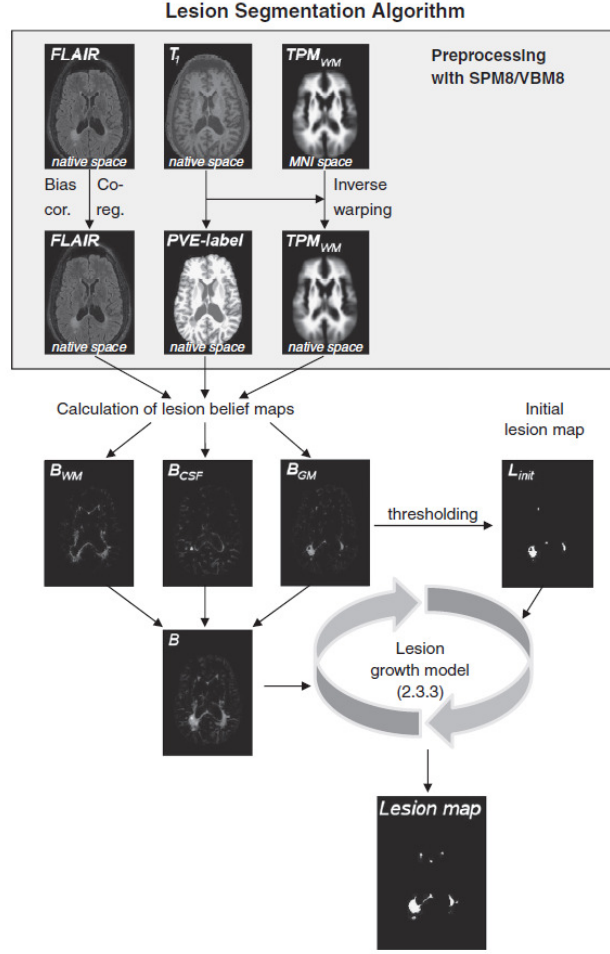


Figure 3.2: The flow diagram of the lesion segmentation algorithm. Image adopted from [218].

based on the information from both T1 and FLAIR images, as shown in Fig. 3.2. The tool calculated the FLAIR intensity distribution for each of the three tissue classes (B_{WM} , B_{CSF} , B_{GM}) to determine outliers, weighted according to the spatial probability of being WM. This resulted in three classes of belief maps summed to generate a single belief map (B). By thresholding the GM lesion belief map with a pre-chosen initial threshold (κ), a conservative binary lesion map was obtained which was subsequently grown along voxels that appeared hyperintense in the FLAIR image. Markov Random Field (MRF) was also integrated into the lesion growing to incorporate the spatial information. Please refer to [218] for complete details. One disadvantage of this unsupervised algorithm is the choice of the initial threshold. Different values yield different segmentation results, which requires sufficient visual inspection of the segmentation.

EM-based algorithm There have been many methods that were developed based on EM algorithm. An EM approach proposed by Wells et al. [269] classified voxels into three classes (or clusters) according to their intensity using a finite Gaussian mixture model with the expectation-maximization (EM) algorithm. The Gaussian mixture model assumed that each class of the model follows a Gaussian distribution and the parameters of each Gaussian can be easily estimated using an EM algorithm. Another approach [76] used multiple Gaussians in the Gaussian mixture model to include partial volumes, which provided a more accurate estimation of the mean and variance of each tissue and, hence, a more accurate segmentation of WM lesions. A more recent EM-based approach proposed by Wang et al. [262] used multi-stage segmentation approach for labeling the WMHs using multi-modal MR images, which first automatically segmented brain tissues on T1 image, and then identified hyperintensities based on FLAIR image, followed by labelling WMHs, cortical infarct and lacunar infarct using T1, T2 and FLAIR images. Such EM-based algorithms commonly model the brain tissues using a Gaussian mixture model, and then estimate the mean and variance of each tissue with EM algorithm. This kind of approaches share a common disadvantage with Bayesian approaches, that is, the assumption for the data distribution probably may not be true in some cases, and hence, satisfactory results can not be produced.

Outlier detection methods There are some other types of methods that are designed not to model the lesions, but to consider them as outliers to the normal appearing brain tissues. Van Leemput et al. [251] employed a weighted EM framework in which voxels far from the model were weighted less in the estimation and considered to be potential lesions. A probabilistic model called Constrained Gaussian Mixture Model (CGMM) [84] was also proposed with lesions identified as outlier Gaussian components. In addition, Weiss et al. [268] proposed to use dictionary learning to learn a sparse representation from pathology-free T1-weighted brain MR scans and then applied this dictionary to sparsely reconstruct brain MR images that contain pathologies, where the lesions were identified using the reconstruction error. The advantages of these approaches are that they avoid the need to model the intensity of the heterogeneous lesions and also provide a more robust estimation in the presence of other tissues or artefacts.

However, probabilistic models in this category have the same disadvantage with EM-based algorithms.

Fuzzy C-means methods Many fuzzy C-means algorithms have also been developed for WMH segmentation. The fuzzy C-means [32] provided a fuzzy clustering of the voxels estimating the center of each class and was less sensitive to the initialisation. Gibson et al. [91] developed a fuzzy-inference approach which included a two-class fuzzy C-means clustering method. Each voxel was tested twice by the fuzzy clustering algorithm, and the result were determined by the consensus of the two segmentations. Additionally, Kawa and Pietka [133] proposed to include spatial information in the clustering using kernel fuzzy C-means combined with fuzzy connectedness theory. However, it requires that the clusters should have a spherical shape for the fuzzy C-means approach to be accurate. The fuzzy C-means can be seen as a special case of the GMM-EM with the assumption that the classes have identical, isotropic covariance matrices [34].

Image synthesis based methods Additionally, several works have also focused on exploiting the fact that WMHs are best observed in FLAIR MR images, while being difficult to identify in T1-weighted MR images. These methods rely on generating a synthetic FLAIR image based on the observed T1-weighted MR image using random forests [277], generative mixture-models [48], support vector regression (SVR) [38] or convolutional neural networks (CNN) [252]. Both synthetic (healthy looking) and real FLAIR (with pathologies) images are then compared to detect any abnormalities. For instance, Bowles et al. [37] proposed to segment WMHs by comparing a real FLAIR image with a subject-specific pathology-free synthetic FLAIR image, which was generated from a T1-weighted image by a modality transformation technique.

Apart from the above mentioned approaches, other method such as lesion-TOADS [226] combines atlas segmentation with statistical intensity modeling to simultaneously segments major brain structures as well as lesions. Some of the probabilistic generative models of the lesion formation for stroke lesion segmentation were also designed, such as [66, 82]. Forbes et al. [82] proposed a Bayesian multi-sequence Markov model for fusing multiple MR sequences to

robustly and accurately segment brain lesions. Derntl et al. [66] proposed to combine standard atlas-based segmentation with a stroke lesion occurrence atlas, in a patient-specific iterative procedure.

An important drawback of all these methods is that they are in fact abnormality detection algorithms and not specifically WMH segmentation or stroke lesion segmentation methods, and hence in principle they detect any pathologies, whether or not a WMH-related pathology. In addition, such unsupervised approaches cannot always produce satisfactory results in subjects with NDs, since WMHs in those subjects are often small, irregular, and heterogeneous within and across subjects [121].

Semi-automatic and semi-supervised segmentation

Several semi-automatic algorithms proposed in the literature for WMH segmentation rely on region growing techniques that require initial seed points to be placed by an operator [264]. For instance, Kawata et al. [134] introduced a region growing method for adaptive selection of segmentation. It segmented WMH regions on the subtraction image between a T1-weighted and FLAIR images using two segmentation methods (a region-growing technique and a level-set method). They were selected for each WMH region based on its image features extracted from initially identified WMH candidates by using a support vector machine. Itti et al. [122] proposed another region growing algorithm that extracts WMHs by propagating seed points into neighboring voxels whose intensity is above an optimised threshold. The process iterates until convergence, i.e. all voxels above the threshold that are connected to the initial seed point had been annotated. Aside from the drawback of requiring per image expert inputs, semi-automatic methods have the additional potential drawback that seeds points could easily be selected in obvious regions, while the biggest challenge of WMH segmentation can arguably be found in more confusing border regions.

The problem of transferring useful knowledge from unlabelled data to a task defined by partially annotated data remains a challenge and an open field of research in its own right. In practice, semi-automatic or semi-supervised WMH segmentation methods, even though they

still require some expert input, tend to underperform when compared to supervised methods, even when the later are trained with only a modest amount of data.

Supervised segmentation

Supervised methods for lesion segmentation have also been well researched. They learn the characteristic features of lesions from training samples that have been manually segmented by an expert. Classical supervised machine learning methods such as k-nearest neighbors (kNN) [12], Bayesian models [170], support vector machines (SVM) [146], and random forests [87] have been well studied in MS and WMH segmentation. We will compare some of them in the following.

Lesion prediction algorithm SPM's LST toolbox also provides a supervised lesion segmentation algorithm, the lesion prediction algorithm (LPA) [217], as an alternative to LGA. LPA is a supervised method which was trained by a logistic regression model with the data of 53 MS patients with severe lesion patterns. As covariates for this model, a similar lesion belief map as for the LGA [218] was used as well as a spatial covariate that took into account voxel specific changes in lesion probability. Parameters of this model fit are used to segment lesions in new images by providing an estimate for the lesion probability of each voxel. Example segmentation results comparing LGA and LPA on three different subjects with NDs are shown in Fig. 3.3.

Supervised inference methods (SVM/Random Forest) Ithapu et al. [121] formulated the hyperintensity detection as a supervised inference problem, and adapted two learning models (SVM and random forest) for this task. Specifically, they extracted texture intensity variation based features from FLAIR images using standard image processing filtering operations. By using the extracted features, a WMH classifier (SVM and random forest) was then constructed from the manual segmentation of FLAIR images. Their proposed method was able to generate a probability estimate of a voxel being WMH or not, thereby discriminating WMH voxels and non-WMH voxels. Extensive simulations implemented in [121] showed that the random forest

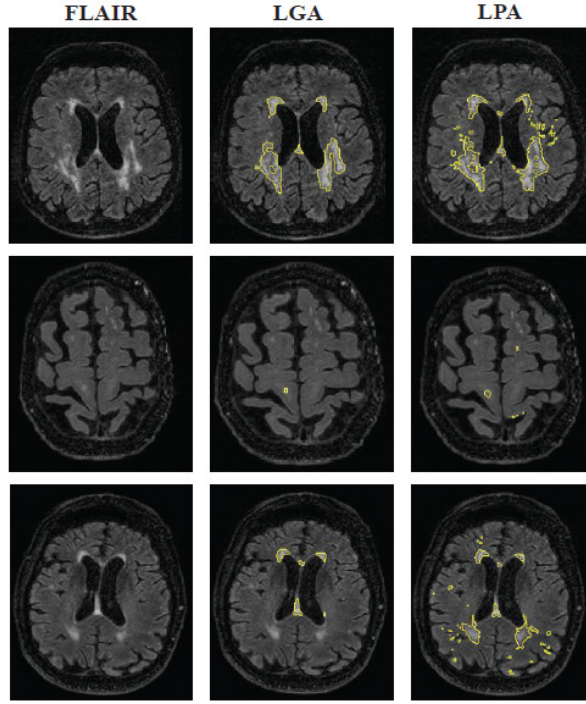


Figure 3.3: Example segmentation results comparing LGA and LPA on subjects with NDs.

based regression worked the best with significant improvement over LGA, the current state-of-the-art unsupervised model. In addition, Maier et al. [167, 168, 169] also proposed to employ SVM and random forest to learn a segmentation function for stroke lesion segmentation.

K-nearest neighbor based algorithm One popular algorithm that used kNN for WMH segmentation is proposed by Anbeek et al. [12]. The algorithm employed information from T1-weighted, inversion recovery, proton density-weighted (PD), T2-weighted and FLAIR scans. It utilised both voxel intensities and spatial information to build a feature space. A learning set is randomly selected from many preclassified voxels. The probability of a voxel being WMH or not was defined as the fraction of hyperintense voxels among its 100 neighbors. Brain intensity abnormality classification algorithm (BIANCA) [97], a fully automated supervised method was also proposed for WMH segmentation based on kNN algorithm. One common limitation for these algorithms is that the segmentation results rely heavily on the selection of the training set. If the training data is not representative of real data, the output results can be unsatisfactory.

Bayesian approaches The Bayesian framework has also been employed in the segmentation of WM lesions. In this context, the training dataset was used to compute the prior probability model of the segmentation, which was then combined with the observed data using Bayes' theorem. An extension of this framework considers that the intensity of the voxels of each tissue follow a Gaussian distribution, and the training data is employed to determine the parameters of the Gaussian distribution for each tissue. Harmouche et al. [104] proposed to model each region of the brain with a different Gaussian distribution to refine the segmentation as well as to incorporate MRF to take advantage of local spatial information. Other methods such as [61, 170] have also been developed recently to classify voxels as normal tissues or WMHs by using Bayesian approaches. Of particular interest and most related to our work, Dalca et al. [59] proposed to use a generative probabilistic model for the differential segmentation of leukoaraiosis and stroke lesions by learning the spatial distribution and intensity profile of each pathology. For these Bayesian approach based methods, they rely on the assumption that the data is generated from a particular distribution. However, this assumption might not hold true in some cases, and therefore leads to only moderately good performance for WMH segmentation.

Many of the above algorithms were originally designed for lesion detection in MS patients, nonetheless, their underlying assumptions allow them to generalize to other types of lesion segmentation. However, in practice, these methods can see a reduction in performance when applied to older subjects, due to the fact that there is an age related decrease in contrast between grey matter (GM) and WM in MR images, and that the boundaries of MS lesions are often less diffuse than those of WMHs [6, 45]. In addition, most of these approaches are not able to discriminate lesions originated from different pathologies. They could fail to differentiate them when different types of lesions coexist and appear similar on one subject.

3.3.2 Deep Learning Methods

More recently, CNNs have been put forward to replace the inference step in many computer vision related tasks [74, 92, 107, 160]. CNNs have been shown to have enough capacity to

model complex nonlinear functions which are capable of performing multi-class classification tasks such as those required for the description and understanding of highly heterogeneous problems.

CNNs have emerged as a powerful alternative for supervised learning on image segmentation tasks, and a variety of methods have been proposed for brain lesion segmentation in recent years [33, 41, 63, 130, 131, 176, 249, 283]. Deep CNNs have the ability to learn discriminant features which may outperform the hand-crafted and pre-defined feature sets. For instance, Brosch et al. [41] proposed a deep convolutional encoder network which combines feature extraction and segmentation prediction on MS lesions. Their work was later extended to a 3D deep encoder network with shortcut connections, which consistently outperformed other methods across a wide range of lesion sizes [40]. The winner of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge [246] developed 3D CNNs to model a voxel-wise classifier which used multi-channel 3D patches of MRI volumes as input. For each ground truth, a CNN was trained and the final segmentation was obtained by combining the probability outputs of these CNNs. Besides, Ghafoorian et al. [88] proposed a CNN architecture that considered multi-scale patches and explicit location features during training, and later the work was extended to consider non-uniform patch sampling [89]. Their best performing architecture shares a similar design with the architecture of DeepMedic proposed by Kamnitsas et al. [130, 131], in which it trained independent paths of convolutional layers for each scale.

Among all these deep learning based medical image segmentation algorithms, DeepMedic [131] and U-net [206] are two approaches that have been widely used for brain lesion segmentation. Details of these two methods will be briefly introduced in the following.

DeepMedic

DeepMedic, proposed by Kamnitsas et al. [131], is a network architecture with two parallel convolutional pathways that process 3D patches at different scales, as shown in Fig. 3.4. In particular, the network consists of a dual path way, 11-layers deep 3D CNNs. Local and larger contextual information is incorporated by introducing the dual pathways, where the

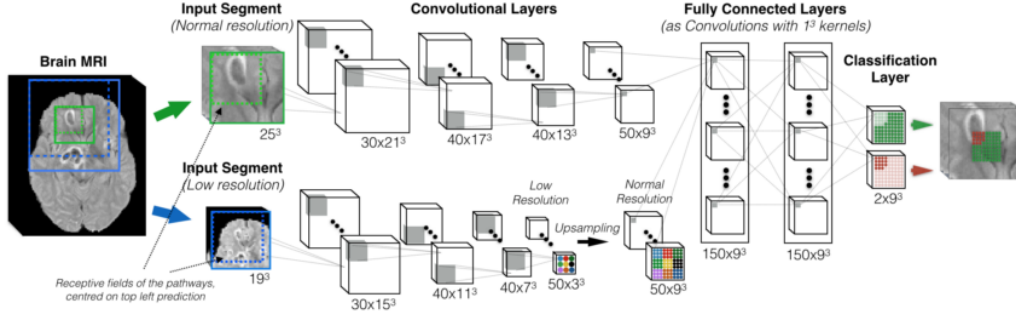


Figure 3.4: Network architecture of DeepMedic for brain tumor segmentation. Image adopted from [131].

detailed local information is captured in the first pathway, while the second pathway performs on the down-sampled images capturing higher level features such as location. In this case, the network is able to process the multi-scale images simultaneously, and the architecture enables the feature extraction independently from multiple scales. This network is followed by a 3D fully connected conditional random field (CRF) as a post-processing step refining the results from CNN to achieve more structured predictions. Although the method was originally proposed for ischemic stroke, tumor and brain injury segmentation on MR images, it can be easily adapted for different tasks using their provided package DeepMedic¹.

Using multi-resolution inputs [88, 89, 131] can increase the field of view with smaller feature maps, while allowing more non-linearities (more layers) to be used at higher resolution, and both of which are desired properties. However, down-sampling patches has the drawback that valuable information is being discarded before any processing is done, and since filters learnt by the first few layers of CNNs tend to be basic feature detectors, e.g. lines or curves, different paths risk capturing redundant information. Furthermore, although convolutions performed in 3D as in [131] intuitively make sense for 3D volumetric images, FLAIR image acquisitions are actually often acquired as 2D images with large slice thickness and then stacked into a 3D volume. Further to this, gold standard annotations, such as those generated by trained radiologists (e.g. WMH delineation or Fazekas scores) are usually derived by assessing images slice by slice. Thus, as pointed out by Ghafoorian et al. [88], 3D convolutions for FLAIR MR image segmentation are in fact less intuitive.

¹<https://github.com/Kamnitsask/deepmedic>

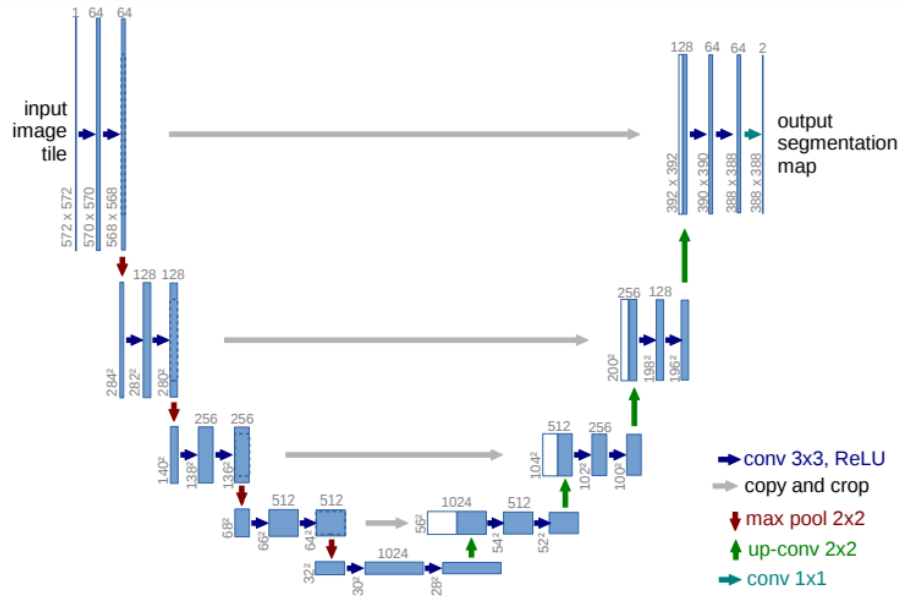


Figure 3.5: Network architecture of U-net for biomedical image segmentation. Image taken from [206].

U-net

Similar to [160], Ronneberger et al. [206] used a U-shaped architecture (U-net) to segment microscopical cell images. The U-net architecture symmetrically combined a contracting and expanding path via feature concatenations, in which up-sampling operations were implemented with trainable kernels (deconvolution or transposed convolutions). In details, as shown in Fig. 3.5, the contracting path follows a standard convolutional network architecture, which consists of a series of 3×3 convolutional kernels, rectified linear units (ReLU), and 2×2 max pooling layers with stride 2 for downsampling. At every level of the upsampling path, it consists of an upsampling and a convolution of the previous feature map, and a concatenation with the corresponding feature map from the contracting path. In this way, the final prediction of segmentation is informed of both higher level features from the expanding path and the lower level features from the contracting path, which allows for a more accurate segmentation performance. This approach has won the ISBI cell tracking challenge in 2015, and has been widely used in various applications, such as cardiac ventricle segmentation [241] and brain tumor segmentation [75].

Table 3.1: Metrics used to evaluate segmentation performance.

Metric	Formula
Dice similarity coefficient (DSC)	$\frac{2 \times TP}{FP + FN + 2 \times TP}$
Accuracy (ACC)	$\frac{TP + TN}{TP + FP + TN + FN}$
Recall	$\frac{TP}{TP + FN}$
Precision	$\frac{TP}{TP + FP}$
Geometric mean (Gmean)	$\sqrt{Precision \times Recall}$

3.4 Evaluation Metrics

Overlap measures

Overlap measures are often used to measure the degree of closeness between the automated segmentation results and the ground truth. In binary classification, there are four basic possible outcomes that give insight into a classifier's performance: true positives (TPs) and true negatives (TNs), where the segmentation is correct, and false positives (FPs) and false negatives (FNs), where segmentation results are not consistent with the ground truth. From these four basic measurements several other measurements can be derived that focus on different aspects of performance.

The most widely used measure for evaluation of segmentation performance is the similarity index or Dice similarity coefficient (DSC) [67]. The value of the index varies between 0 and 1 (perfect segmentation), with 0.7 normally considered as a good segmentation. The formulation of this measure is shown in Table 3.1. A schematic illustration of measuring the segmentation errors for calculating the DSC is shown in Fig. 3.6. In the context of WMH segmentation, TNs are much more than TPs, which leads to the lack of information about under segmentation when using DSC metric. Other evaluation measures include recall, accuracy, precision, and geometric mean (Gmean). The definitions of these measures are given in Table 3.1. The values of these similarity metrics vary between 0 and 1, with higher values indicating better performance.

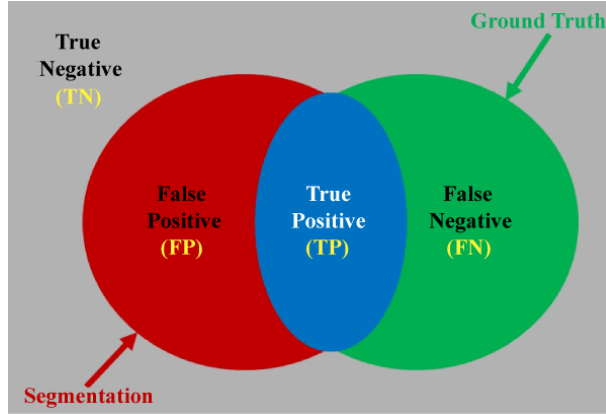


Figure 3.6: Schematic illustration of measuring the segmentation errors for calculating the Dice similarity coefficient (DSC). Image adopted from [174].

Lesion-based measures

Total lesion load (TLL) is often used as a biomarker in clinical trials, and has also been employed to evaluate the performance of lesion segmentation methods, such as the difference between automatic segmentation and manual segmentation [157]. Bland and Altman analysis [35] is often employed to assess the comparability between automatic segmentation volumes and manual segmentation volumes by studying the mean difference and constructing limits of agreement. The coefficient of determination R^2 measures the regression performance between these two volumes, where a linear regression model is used with the ideal trend $f(x) = 1.0x + 0.0$. An R^2 of 1 indicates that the regression line perfectly fits the data. Additionally, clinical validation, such as correlation between WMH volumes and Fazekas scores is used to show the consistency between automatic segmentation results and clinical scores. The main limitation of using lesion volume as a measure is that there is no information regarding the overlap of the segmentation. In the extreme case, the automatic segmentation method could obtain the same TLL as the true segmentation but have no voxels in common. Number of lesions is also often used in diagnosis and as an output measure in clinical trials, such as comparing the number of lesions obtained in automatic segmentation with that of manual segmentation [211, 234]. However, such methods cannot provide any information of the accuracy of lesion boundary, as well as the over- and under-segmentation.

3.5 Conclusions

This chapter has provided an overview of WMH segmentation. Background and challenges of WMH segmentation task has been introduced. Recent research and related works on WMH segmentation and more general lesion segmentation have also been presented, including both traditional methods and deep learning methods. Besides, general preprocessing steps before lesion segmentation are given, as well as the evaluation metrics commonly used for measuring segmentation performance. The following chapters in this part will introduce two novel methods researching on this aspect.

Chapter 4

Large Margin Algorithm for WMH Segmentation

This chapter is based on:

- **C. Qin**, R. Guerrero, C. Bowles, C. Ledig, P. Scheltens, F. Barkhof, H. Rhodius-Meester, B. Tijms, A. W. Lemstra, W. M. van der Flier, B. Glocker, D. Rueckert. A Semi-supervised Large Margin Algorithm for White Matter Hyperintensity Segmentation. International Workshop on Machine Learning in Medical Imaging. Springer, 2016: 104-112.
- **C. Qin**, R. Guerrero, C. Bowles, L. Chen, D. A. Dickie, M.C. Valdés-Hernández, J. Wardlaw, D. Rueckert. A Large Margin Algorithm for Automated Segmentation of White Matter Hyperintensity. Pattern Recognition, 2018, 77: 150-159.

4.1 Introduction

WMHs are areas of the brain in cerebral WM that appear bright on T2-weighted FLAIR MR images due to localised, pathological changes in tissue composition [266], as described in Chapter 3. Precise detection and quantification of WMHs is of great interest in studies of neurological and vascular disorders. However, manual annotation is time consuming and can

introduce inter- and intra-observer variability. Thus, it is highly desirable to develop reliable and automated WMH segmentation methods.

In this chapter, a novel approach for WMH segmentation is proposed which includes a supervised large margin algorithm (SLM) followed by a semi-supervised large margin algorithm (SSLM) in the framework. Specifically, both the supervised and semi-supervised large margin approaches optimise a kernel based max-margin objective function formulated to maximise the margin between the two class data (inliers and outliers) while enforcing the consistency between the predictions and the labels. Compared with the proposed SLM method where the classifier is trained given the ground truth labels, the proposed SSLM model learns a large margin classifier and a label assignment by iteratively updating the classifier and the label indicator. The key concept behind the proposed SSLM approach is to tackle the uncertainty of the unlabelled input data with the help of a small proportion of labels, and to discover outliers (WMH) by training a classifier which maximises the average margins between the estimated inliers (normal tissues) and outliers. Instead of assuming that data is generated from a particular distribution, as most outlier detection methods do [251, 276] which may not hold true for WMH segmentation, our method assumes that neighboring patches in feature space tend to have consistent classifications.

In particular, the proposed approach consists of a two-step segmentation framework, where a general SLM classifier is first trained across different subjects in the training data followed by learning a SSLM classifier for each individual test subject to obtain the hard labels based on the predictions from SLM. The labelled information for SSLM is determined by thresholding the prediction map of SLM while the precision of such labelled samples and their influence on the guidance for SSLM learning are controlled by their corresponding confidence values. In such case, no ground truth labels are required for learning SSLM and it works in an unsupervised way in practice. We name this framework as Combined Supervised and Semi-Supervised Large Margin (CS³LM) method. In this way, the proposed method is fully automatic and can utilise both common information across subjects and individual information from the target subject. Quantitative results indicate the competitiveness and effectiveness of the proposed algorithms against other popular methods for WMH segmentation.

4.2 Background and Notation

Let $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$ denote a set of n unlabelled input samples, and y_i represent the corresponding unknown soft label that assigns a positive value (c^+) for normal samples and a negative value (c^-) for outliers. Additionally, let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) of the function: $f(\mathbf{x}) = \sum_{i=1}^n \kappa(\mathbf{x}, \mathbf{x}_i) \alpha_i$, with associated kernel κ as the functional base and the expansion coefficient α . Unsupervised one-class learning (UOCL) [159] is an unsupervised algorithm that uses a self-guided labelling procedure to discover potential outliers in the data, which has been shown to be robust to a high outlier proportion. This method aims to separate inliers from outliers by training a large margin classifier, which is obtained from minimising the following objective function:

$$\begin{aligned} \min_{f \in \mathcal{H}, \{y_i\}} \quad & \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 + \gamma_1 \|f\|_{\mathcal{M}}^2 - \frac{2\gamma_2}{n_+} \sum_{i, y_i > 0} f(\mathbf{x}_i) \\ \text{s.t.} \quad & y_i \in \{c^+, c^-\}, \forall i \in [1 : n], \quad 0 < n_+ = |\{i | y_i > 0\}| < n. \end{aligned} \quad (4.1)$$

The first term in Equation (4.1) uses the squared loss to make the classification function consistent with the label assignment. In order to remove the influence of the varying $\|\mathbf{y}\|^2 = \sum_{i=1}^n y_i^2$ on optimisation of Equation (4.1), the values (c^+, c^-) of soft labels are designed such as $(\sqrt{\frac{n-n_+}{n_+}}, -\sqrt{\frac{n_+}{n-n_+}})$, so that $\|\mathbf{y}\|^2$ is constant, in which n is the total number of samples and n_+ stands for the number of samples with positive labels. The second term of Equation (4.1) is a manifold regulariser [30], which endows f with the smoothness along the intrinsic manifold structure $\mathcal{M}$ underlying the data. This term is constructed by using a k NN graph with affinity matrix defined as

$$W_{i,j} = \begin{cases} \exp\left(-\frac{\mathcal{D}(x_i, x_j)}{\varepsilon^2}\right), & i \in \mathcal{N}_j \text{ or } j \in \mathcal{N}_i, \\ 0, & \text{otherwise,} \end{cases} \quad (4.2)$$

where $\mathcal{D}(\cdot)$ is a Euclidean distance measure, the set $\mathcal{N}_i \subset [1 : n]$ contains the indices of k nearest neighbors of x_i in $\mathcal{X}$, and ε is the bandwidth parameter. Then the manifold regulariser can be written as:

$$\|f\|_{\mathcal{M}}^2 = \frac{1}{2} \sum_{i,j=1}^n (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2 W_{ij} = \mathbf{f}^T \mathbf{L} \mathbf{f}. \quad (4.3)$$

Here, $\mathbf{f} = \begin{bmatrix} f(\mathbf{x}_1), & \dots, & f(\mathbf{x}_n) \end{bmatrix}^T \in \mathbb{R}^n$, T is the transpose operator, and the graph Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with diagonal elements being $D_{ii} = \sum_{j=1}^n W_{ij}$. The third term of Equation (4.1) represents the margin averaged over the positive samples, which aims to push the majority of the inliers as far away as possible from the decision boundary $f(\mathbf{x}) = 0$ while suppressing the bias caused by the dubious outliers. The importance of all three terms are balanced by the trade-off parameters γ_1 and γ_2 . A Gaussian kernel $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2)$ is used in the classification function.

For concise notation, the vectorial kernel mapping can be defined as

$$\mathbf{k}(\mathbf{x}) = \begin{bmatrix} \kappa(\mathbf{x}_1, \mathbf{x}), & \dots, & \kappa(\mathbf{x}_n, \mathbf{x}) \end{bmatrix}^T, \quad (4.4)$$

and the kernel matrix is $\mathbf{K} = [\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{1 \leq i, j \leq n}$, so the target function can be expressed as $f(\mathbf{x}) = \alpha^T \mathbf{k}(\mathbf{x})$ and $\mathbf{f} = \mathbf{K}\alpha$, in which $\alpha = \begin{bmatrix} \alpha_1, & \dots, & \alpha_n \end{bmatrix}^T \in \mathbb{R}^n$. Thus, by incorporating Eq. (4.3) and ignoring the constant term $\|\mathbf{y}\|^2$, the objective function can be rewritten as follows:

$$\begin{aligned} \min_{\alpha, \mathbf{y}} \quad & \alpha^T \mathbf{K}(\mathbf{I} + \gamma_1 \mathbf{L}) \mathbf{K} \alpha - 2\alpha^T \mathbf{K}(\mathbf{y} + \tilde{\mathbf{y}}) \\ \text{s.t.} \quad & \mathbf{y} \in \{c^+, c^-\}^{n \times 1}, \quad \tilde{y}_i = \begin{cases} \frac{\gamma_2}{\|\mathbf{y}\|_+}, & y_i = c^+, \\ 0, & y_i = c^-. \end{cases} \end{aligned} \quad (4.5)$$

in which $\|\mathbf{y}\|_+$ stands for the number of positive elements in vector $\mathbf{y}$, and $\mathbf{I} \in \mathbb{R}^{n \times n}$ stands for the identity matrix. This minimisation problem is solved by an alternating optimisation scheme, with the continuous function f and discrete label assignment y_i being optimised iteratively. For more details, please refer to [159]. The UOCL method has several advantages that can be beneficial for WMH segmentation. Firstly, unlike most of the other outlier detection methods, it makes no assumption of the data distribution. Secondly, it works under a self-guided mechanism learning a large margin classifier that directly indicates inliers and outliers, which is advantageous for the SSLM proposed in Section 3.1.2. In addition, it has been shown to be robust to high outlier proportion, which is a highly desirable trait in WMH segmentation.

4.3 Supervised and Semi-supervised Large Margin Algorithms

When it comes to WMH segmentation, the classification results of UOCL are not always satisfactory. Since outliers can originate from low-density samples and be later separated from high-density regions without guidance from labelled information, the UOCL method can produce many false positives (FPs) when segmenting WMH. In particular, intensity edges and partial volumes can be identified as outliers. To address this problem, we developed a new SLM algorithm and SSLM algorithm for WMH segmentation. Different from the SSLM method we proposed in [196], here we learn a general SLM classifier across subjects (MR images) to provide a rough lesion segmentation map, and then SSLM is employed for the label refinement on per test subject. This allows for the information from both the training images (SLM) and the target image (SSLM) to guide the classification. We will introduce the two methods in detail in the following subsections.

4.3.1 Supervised Large Margin Algorithm

SLM learning model

We adopt the same form of classifier as in UOCL, which is $f(\mathbf{x}) = \sum_{i=1}^n \kappa(\mathbf{x}, \mathbf{x}_i) \alpha_i$. However, in order to apply the trained classifier in test data, in contrast to UOCL which learns labels and classifiers simultaneously, the proposed SLM needs a fixed set of reference samples to compute the $\kappa(\mathbf{x}, \mathbf{x}_i)$ during inference. Therefore, following the notation in Section 4.2, we define a set of learnt reference samples $\mathcal{Z} = \{\mathbf{z}_j \in \mathbb{R}^d\}_{j=1}^m$ for the classifiers so that the classification function can be formulated as $f(\mathbf{x}) = \sum_{j=1}^m \kappa(\mathbf{x}, \mathbf{z}_j) \alpha_j$, in which α_j and $\mathbf{z}_j$ are parameters needed to be learnt, and m is the number of kernel functions and also the the number of reference samples which should be pre-set.

Now we establish our SLM model as minimising the following objective:

$$\begin{aligned} \min_{f \in \mathcal{H}} \quad & \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 - \frac{2\gamma_1}{n_+} \sum_{i, y_i > 0} f(\mathbf{x}_i) + \frac{2\gamma_2}{n_-} \sum_{i, y_i < 0} f(\mathbf{x}_i) \\ \text{s.t.} \quad & y_i \in \{c^+, c^-\}, \forall i \in [1 : n], \quad n_+ = |\{i | y_i > 0\}|, n_- = |\{i | y_i < 0\}|, \end{aligned} \quad (4.6)$$

where variables γ_1 and γ_2 are trade-off parameters controlling the model, and n_+ and n_- are numbers of positive and negative samples respectively during learning, and y_i are known labels in this case. Different from UOCL formulation, motivated by [271], here we introduce a new term $\sum_{i, y_i < 0} f(\mathbf{x}_i)/n_-$ into the objective function given by Eq. (4.6), which aims to maximise the average margin between the outliers and the decision boundary. The last two terms in objective function (4.6) work together to push the positive samples and outliers far away from the decision boundary, thus enabling these two groups of data to be far away from each other. In addition, as learnt from experiments, including $\|f\|_{\mathcal{M}}^2$ in Eq. (4.1) does not affect the performance significantly. Therefore, we remove this term in the new formulation for simplicity. As in the supervised training, n_+ and n_- are fixed during the learning, thus we are able to replace $2\gamma_1/n_+$ and $2\gamma_2/n_-$ with γ_1 and γ_2 , which means that in the supervised setting, maximising the summed margin is equivalent to maximising the averaged margin. Therefore, the proposed objective function in vector-matrix form can be written as follows:

$$\begin{aligned} Q(\alpha, \mathbf{z}) &:= \min_{\alpha, \mathbf{z}} \alpha^T \mathbf{K}^T \mathbf{K} \alpha - 2\alpha^T \mathbf{K}^T (\mathbf{y} + \tilde{\mathbf{y}}) \\ \text{s.t.} \quad & \mathbf{y} \in \{c^+, c^-\}^{n \times 1}, \quad \tilde{\mathbf{y}} = \begin{cases} \gamma_1, & y_i = c^+, \\ \gamma_2, & y_i = c^-. \end{cases} \end{aligned} \quad (4.7)$$

Different from Eq. (4.5), here $\mathbf{K} = [\kappa(\mathbf{x}_i, \mathbf{z}_j)]_{1 \leq i \leq n, 1 \leq j \leq m} \in \mathbb{R}^{n \times m}$. In the SLM algorithm, instead of learning the labels $\mathbf{y}$, we need to learn the representation $\mathbf{z}_j$ and coefficient α for the classifier $f(\mathbf{x}) = \sum_{j=1}^m \kappa(\mathbf{x}, \mathbf{z}_j) \alpha_j$.

SLM Optimisation

Here we use gradient descent to minimise the objective function (4.7) in order to learn the parameters $\alpha \in \mathbb{R}^m$ and $\mathbf{z}_j \in \mathbb{R}^{m \times d}, j = 1, \dots, m$. The gradients of the objective function with

respect to α and $\mathbf{z}_j$ are as follows:

$$\frac{\delta Q}{\delta \alpha} = 2(\mathbf{K}^T \mathbf{K} \alpha - \mathbf{K}^T \mathbf{y} - \mathbf{K}^T \tilde{\mathbf{y}}), \quad (4.8)$$

$$\frac{\delta Q}{\delta \mathbf{z}_j} = 2\alpha_j [\mathbf{K}_{ij}(\mathbf{x}_i - \mathbf{z}_j)]_{1 \leq i \leq n} (\mathbf{K} \alpha - \mathbf{y} - \tilde{\mathbf{y}}). \quad (4.9)$$

Here $[\mathbf{K}_{ij}(\mathbf{x}_i - \mathbf{z}_j)]_{1 \leq i \leq n} = [\mathbf{K}_{1j}(\mathbf{x}_1 - \mathbf{z}_j), \dots, \mathbf{K}_{ij}(\mathbf{x}_i - \mathbf{z}_j), \dots, \mathbf{K}_{nj}(\mathbf{x}_n - \mathbf{z}_j)]$. By minimising the objective function using the conjugate gradient method with the obtained gradients, we are able to learn a supervised classifier for WMH segmentation.

4.3.2 Semi-supervised Large Margin Algorithm

Though SLM can learn a binary classifier based on features extracted from training images, the classifier is not informed of the feature characteristics of the target image. Even though test images should be intensity normalised in a preprocessing step, there still exists heterogeneity within and across subjects with WMHs. In the following, we propose a semi-supervised large margin algorithm to learn the labels for samples with the most uncertainty on an individual subject, which can be viewed as a label refinement method to tackle the heterogeneity within subject that lies in the same data distribution where the model is trained. The labelled information for the SSLM model is determined based on the predictions of SLM. Note that we name the method as a semi-supervised approach because of its formulation, however, no ground truth labels are needed in the training and the segmentation method works in a fully automated way.

SSLM Learning Model

Following the notation defined in Section 4.2, we define L as the labelled data set and U as the unlabelled data set, which, in the WMH segmentation case, represent sets of voxels with known and unknown labels respectively. In experiments, the labelled data set is determined by the prediction from the SLM algorithm with a threshold applied. Based on the formulation of SLM, we incorporate the labelled information and the objective function of our proposed

semi-supervised model is thus formulated as:

$$\begin{aligned}
& \min_{f \in \mathcal{H}, \{y_i\}} \sum_{\mathbf{x}_i \in U} (f(\mathbf{x}_i) - y_i)^2 + \sum_{\mathbf{x}_j \in L} \lambda_j (f(\mathbf{x}_j) - y_j)^2 \\
& \quad - \frac{2\gamma_1}{n_+} \sum_{k, y_k > 0} f(\mathbf{x}_k) + \frac{2\gamma_2}{n_-} \sum_{k, y_k < 0} f(\mathbf{x}_k), \\
& \text{s.t. } y_i \in \{c^+, c^-\}, \quad n_+ = |\{i | y_i > 0\}|, \quad n_- = |\{i | y_i < 0\}|,
\end{aligned} \tag{4.10}$$

where $\mathbf{x}_k \in L \cup U$, and variable λ_j is a trade-off parameter controlling the influence of labelled data on the model, which in this case is dependent on the confidence of the SLM prediction. To make full use of the limited amount of labelled information and to enable the classification to be informed by the available labels, in this model, we introduce a new term $\sum_{\mathbf{x}_j \in L} \lambda_j (f(\mathbf{x}_j) - y_j)^2$ that represents the squared loss for labelled data, thereby allowing it to better discriminate between inliers and outliers. The classification function is of the same form as in SLM, which is $f(\mathbf{x}) = \sum_{j=1}^m \kappa(\mathbf{x}, \mathbf{z}_j) \alpha_j$. Similarly, the objective function can be rewritten in a vector-matrix form as follows:

$$\begin{aligned}
& \min_{\alpha, \mathbf{z}, \mathbf{y}} \alpha^T \mathbf{K}^T \mathbf{A} \mathbf{K} \alpha - 2\alpha^T \mathbf{K}^T \mathbf{A} \mathbf{y} + \mathbf{y}^T \mathbf{A} \mathbf{y} - 2\alpha^T \mathbf{K}^T \tilde{\mathbf{y}} \\
& \text{s.t. } \mathbf{y} \in \{c^+, c^-\}^{n \times 1}, \quad \mathbf{A} = \text{diag}(1, \dots, 1, \underbrace{\lambda_1, \dots, \lambda_j}_{\mathbf{x}_j \in L}, 1, \dots, 1), \\
& \quad \tilde{y}_i = \begin{cases} \frac{\gamma_1}{\|\mathbf{y}\|_+}, & y_i = c^+, \\ -\frac{\gamma_2}{\|\mathbf{y}\|_-}, & y_i = c^-. \end{cases}
\end{aligned} \tag{4.11}$$

Here, $\|\mathbf{y}\|_+ = n_+$ and $\|\mathbf{y}\|_- = n_-$ respectively stand for the number of positive elements and negative elements in vector $\mathbf{y}$. In the proposed method, the same soft label assignment for (c^+, c^-) as in [159], i.e. $(\sqrt{\frac{n_-}{n_+}}, -\sqrt{\frac{n_+}{n_-}})$ was adopted. Note that different from our previously proposed approach [196], here the SSLM method needs to learn an additional parameter $\mathbf{z}$ inherited from SLM. Compared with UOCL method which learns parameters $(\alpha, \mathbf{y})$, here the proposed SSLM needs to learn $(\alpha, \mathbf{z})$ and $\mathbf{y}$ iteratively.

SSLM Optimisation

Solving the proposed model involves a mixed optimisation of the classifier (including a continuous variable α and continuous representations $\mathbf{z}$) and a discrete variable $\mathbf{y}$. One key observation is that if one of the two components is fixed, the optimisation problem becomes easy to solve. Here, similar to an expectation-maximisation (EM) framework, we propose to alternately optimise the classifier $(\alpha, \mathbf{z})$ and $\mathbf{y}$ via iterative updates.

First, for a given label indicator $\mathbf{y}$, computing the optimal α and $\mathbf{z}$ is equivalent to minimisation of the following sub-problem:

$$\min_{\alpha, \mathbf{z}} Q(\alpha, \mathbf{z}) := \alpha^T \mathbf{K}^T \mathbf{\Lambda} \mathbf{K} \alpha - 2\alpha^T \mathbf{K}^T \mathbf{\Lambda} \mathbf{y} - 2\alpha^T \mathbf{K}^T \tilde{\mathbf{y}}. \quad (4.12)$$

The gradient of the objective function $Q(\alpha, \mathbf{z})$ in Eq. (4.12) with respect to α and $\mathbf{z}_j$ is

$$\frac{\delta Q}{\delta \alpha} = 2(\mathbf{K}^T \mathbf{\Lambda} \mathbf{K} \alpha - \mathbf{K}^T \mathbf{\Lambda} \mathbf{y} - \mathbf{K}^T \tilde{\mathbf{y}}), \quad (4.13)$$

$$\frac{\delta Q}{\delta \mathbf{z}_j} = 2\alpha_j [\mathbf{K}_{ij}(\mathbf{x}_i - \mathbf{z}_j)]_{1 \leq i \leq n} (\mathbf{\Lambda} \mathbf{K} \alpha - \mathbf{\Lambda} \mathbf{y} - \tilde{\mathbf{y}}). \quad (4.14)$$

By using the gradient, Eq. (4.12) can be efficiently solved by the conjugate gradient descent method.

When α and $\mathbf{z}_j$ are fixed, we need to deal with the $\mathbf{y}$ -subproblem with objective function $H(\mathbf{y})$, that is

$$\begin{aligned} \max_{\mathbf{y}} H(\mathbf{y}) &:= 2\alpha^T \mathbf{K}(\mathbf{\Lambda} \mathbf{y} + \tilde{\mathbf{y}}) - \mathbf{y}^T \mathbf{\Lambda} \mathbf{y} \\ s.t. \quad \mathbf{y} &\in \{c^+, c^-\}^{n \times 1}, \quad \tilde{y}_i = \begin{cases} \frac{\gamma_2}{\|\mathbf{y}\|_+}, & y_i = c^+, \\ -\frac{\gamma_3}{\|\mathbf{y}\|_-}, & y_i = c^-. \end{cases} \end{aligned} \quad (4.15)$$

Here, a simpler case is shown to solve this discrete optimisation problem. If an integer $l = \|\mathbf{y}\|_+$ is given, then $\mathbf{y}^T \mathbf{\Lambda} \mathbf{y}$ and the soft label assignment for labelled data remain the same regardless of the label assignment for unlabelled data. Thus this problem reduces to the same one as in UOCL, i.e., to maximise $(\mathbf{K} \alpha)^T (\mathbf{y} + \tilde{\mathbf{y}})$ in the unlabelled data set. It has been shown in [159] that an optimal solution satisfies $y_i > 0$ if and only if f_i is among those largest elements of

Algorithm 1: **SSLM**

Input: Input samples $\mathcal{X}$, kernel matrix $\mathbf{K}$, model parameters $\lambda, \gamma_1, \gamma_2 > 0$, m , $\mathbf{\Lambda}$ and *maxiter*

Initialisation

$\alpha_0 = 1/\sqrt{m}$ or α_{SLM} , $\mathbf{z}_0 = \text{rand}(m, d)$ or $\mathbf{z}_{\text{SLM}}$,
 $l_0 = \arg \max_l H(q(\mathbf{K}\alpha_0, l))$, $\mathbf{y}_0 = q(\mathbf{K}\alpha_0, l_0)$, $\tilde{\mathbf{y}}_0 = h(l_0, \mathbf{y}_0)$, $t = 0$;

repeat

Update α_{t+1} and $\mathbf{z}_{t+1}$ by optimising function (4.12) using conjugate gradient descent method;

Update l_{t+1} : $l_{t+1} = \arg \max_l H(q(\mathbf{K}\alpha_{t+1}, l))$;

Update $\mathbf{y}_{t+1}$ and $\tilde{\mathbf{y}}_{t+1}$: $\mathbf{y}_{t+1} = q(\mathbf{K}\alpha_{t+1}, l_{t+1})$, $\tilde{\mathbf{y}}_{t+1} = h(l_{t+1}, \mathbf{y}_{t+1})$;
 $t = t + 1$;

until convergence or $t > \text{maxiter}$

Output: classifier coefficients $\alpha^* = \alpha_t$ and $\mathbf{z}^* = \mathbf{z}_t$, and the soft label assignment $\mathbf{y}^* = \mathbf{y}_t$.

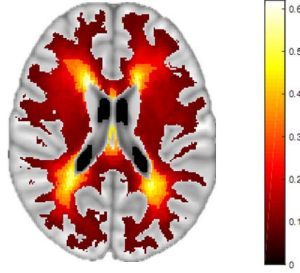


Figure 4.1: The lesion atlas constructed from a different group of subjects.

f. One optimal solution to the Eq. (4.15) can be simply obtained by sorting $\mathbf{f}$ for unlabelled data in a descending order. Then $y_i > 0$ is assigned to samples before and including the l_U -th element, while $y_i < 0$ to those after the l_U -th element. Here $l_U = l - l_L$, where l_U and l_L stand for the number of positive samples in the unlabelled and labelled data sets respectively, with l_L a fixed number. Therefore, the solution to the subproblem given by Eq. (4.15) can be expressed as $\mathbf{y}^*(\alpha) = q(\mathbf{K}\alpha, l^*(\alpha))$, in which $l^*(\alpha) = \arg \max_l H(q(\mathbf{K}\alpha, l))$ and $q(\cdot)$ denotes the function that has the above detailed optimisation process. For simplicity, we further define $\tilde{\mathbf{y}}$ as a function of l and $\mathbf{y}$, i.e., $\tilde{\mathbf{y}} = h(l, \mathbf{y})$, in which $h(\cdot)$ stands for the function that derives $\tilde{\mathbf{y}}$ from $\mathbf{y}$ as detailed in Eq. (4.15). A summary of this method is shown in Algorithm 1, in which α_{SLM} and $\mathbf{z}_{\text{SLM}}$ are values learnt from the SLM classifier.

4.4 Experiments

4.4.1 Data

Data used in the preparation of this work consisted of T1 and FLAIR MR images from 88 subjects with WMHs without confounding radiological evidence of recent or old strokes. All image data was acquired at the Brain Research Imaging Centre of Edinburgh [1] on a GE Signa Horizon HDx 1.5T clinical scanner (General Electric, Milwaukee, WI), equipped with a self-shielding gradient set and manufacturer-supplied eight-channel phased-array head coil. More details can be found in [247]. Since no gold standard for segmentation of WMHs exists, we compared our algorithm with semi-automated computational processing with expert visual correction, in which WMH were extracted following the procedure described in [110, 247].

The WM lesion atlas used in this work, shown in Fig. 4.1, was built on a separate group of data consisting of 277 MR FLAIR images from subjects with cerebral WMHs. The atlas gives a 3D population-based WMH occurrence probabilistic distribution. Details of the atlas construction can be found in [50]. Here the atlas is used as a mask to select the region of interest on the images in the pre-processing step.

4.4.2 MR image pre-processing

All images were coregistered to FLAIR space using FSL-FLIRT [125]. T1 images were segmented using an automated brain segmentation tool [150], from which brain, WM, GM and ventricle probability maps were obtained. A region of interest (ROI) was determined by registering the lesion atlas to the FLAIR space, and we assume that all WMH lie on the registered lesion atlas region in order to reduce the input sample size. Additionally, FLAIR images were intensity normalised as in [119]. For each voxel in the ROI, a feature vector was constructed by the intensity of the voxel, mean and maximum intensities of a 7×7 neighborhood patch from the FLAIR image, as well as the probability of being WM, GM, and lesion. Here we used 2D patches, as FLAIR MR images are commonly acquired using 2D multi-slice acquisitions with

large slice thickness.

4.4.3 Experimental settings

In our experiments, we compared the proposed algorithm with the widely used lesion growth algorithm (LGA) and lesion prediction algorithm (LPA) as implemented in the LST toolbox version 2.0.15 [2]. LPA [217] is a supervised method which was trained using a logistic regression model on the data of 53 MS patients with severe lesion patterns. LGA [218] is an unsupervised method which segments lesions using a combination of FLAIR and T1 images. Through grid search, the optimal κ for LGA was found to be 0.1, and the optimal threshold t used for binarizing the probabilistic segmentations of LPA was 0.2. This differs from the author's suggested threshold of 0.5. We use a Radial Basis Function (RBF) kernel for training a SVM as a comparison. For the combined supervised and semi-supervised large margin (CS³LM) approach, a threshold of 0.7/0.3 is used to threshold the prediction map f of SLM to determine the labelled normal/WMH data, and λ_j is defined as $10 \times f$ when $f > 0.7$, and $10 \times (1 - f)$ when $f < 0.3$, in which case we only refine the label assignments for samples with most uncertainty. The value of parameters m , γ_1 and γ_2 in SLM and SSLM are determined based on the experiments on the validation set. In the following experiments, SLM refers to the supervised large margin method only with a threshold of 0.5 for the prediction, while CS³LM stands for the combined supervised and semi-supervised framework for WMH segmentation.

All these methods were implemented in MATLAB 2015a on a PC with 16GB RAM, 3.60GHz CPU. For SLM, the computation time for each brain scan with volume size of $256 \times 256 \times 30$ is around 0.03s. For SSLM, the computation time for each brain volume is on average around 16s (average convergence steps: 8 iterations). In addition, in experiments, we performed a two-fold validation to evaluate the model's performance. Data was first randomly split into 38/38/12 for training/testing/validation, and subsequently, the training and testing dataset were swapped to ensure that there is no overlap between the training data used in the two experiments. Results reported were averaged values of the two-fold experiments. Evaluation metrics introduced in Chapter 3 were employed to compare the performance.

Table 4.1: Analysis of parameters γ_1 and γ_2 on validation set. Reported numbers are DSC score.

$\gamma_2 \backslash \gamma_1$	0	0.01	0.1	1
0	0.6372	0.6391	0.6344	0.5536
0.01	0.6401	0.6408	0.6361	0.5569
0.1	0.6405	0.6428	0.6411	0.5681
1	0.5629	0.5709	0.5984	0.6336

4.5 Results

4.5.1 Parameter analysis

In this section, we will analyse the effects of parameters m , γ_1 and γ_2 on the model's performance: m is the number of kernel functions for the classifier which determines the model capacity, and γ_1 , γ_2 are trade-off parameters controlling the weights of the terms of the model. As observed from experiments, the effects of these two sets of parameters on the model's performance are relatively independent, thus we will analyse their influence separately. In addition, effect of different initialisation strategies for SSLM optimisation is also analysed in this section.

Effect of the trade-off parameters γ_1 and γ_2 : Table 4.1 shows the experimental results of varying γ_1 and γ_2 , where the performance is evaluated using DSC on the validation data set. In this experiment, m is set to be 50, which gave reasonable results in preliminary experiments. The range of γ_1 and γ_2 is set to be within $[0,1]$. From Table 4.1, the benefit of introducing the max-margin terms can be seen. This justifies the formulation of the large margin classifier. Furthermore, it can be observed that the performance dropped greatly if γ_1 or γ_2 reached 1, as this can lead to the model's bias towards one of the classes, pushing the majority of the samples to the class with larger trade-off parameter. Also, it seems that γ_2 has more influence on the model's performance than γ_1 , where the best performance is achieved with $(\gamma_1, \gamma_2) = (0.01, 0.1)$. This can be explained by the highly imbalanced data in WMH segmentation, where the amount of WMH is less than that of normal tissue. Therefore, a higher γ_2 is beneficial for the classifier to identify the WMH.

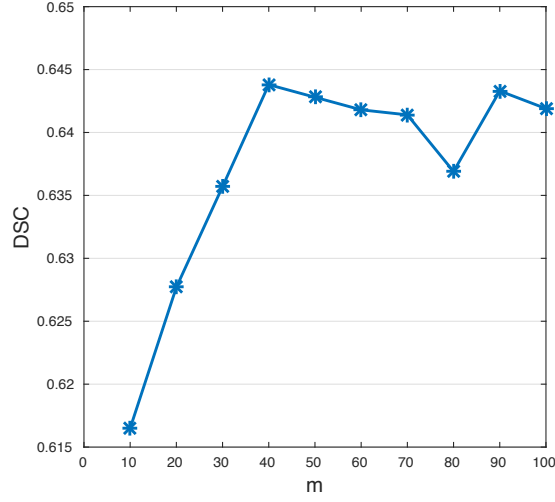


Figure 4.2: The model's performance influenced by the value of m over the validation set.

Effect of the parameter m : Fig. 4.2 shows the DSC scores on the validation set with varying m . In the figure, the x -axis denotes the different m values from 10 to 100 with 10 intervals and y -axis denotes the corresponding DSC value. Here γ_1 and γ_2 are set to be 0.01 and 0.1 respectively as examined in Table 4.1. We can observe from Fig. 4.2 that when m is small, the performance is limited by the model's capacity, and as m increases, the DSC also increases due to the model's higher expressive power. When m reaches a certain value, 40 in this case, the model's performance stays steady though fluctuates slightly. Therefore, we choose $m = 40$ as the model parameter for the following experiments, which gives a good performance while keeping the fast training speed due to the small number of parameters.

Effect of SLM initialisation for SSLM: In Algorithm 1, we have shown that SSLM can be initialised both in a completely random fashion or with initial SLM training. Here we examine the effectiveness of SLM initialisation against random initialisation by computing the DSC for each iteration. The comparison results are shown in Fig. 4.3. We can see that with SLM initialisation, SSLM converges faster and also achieves a higher DSC than the one with random initialisation, which suggests that different initialisation strategies enable the algorithm to converge to a different local optimum as the problem formulation is non-convex. In the following experiments, we will use SLM initialisation for the implementation of SSLM.

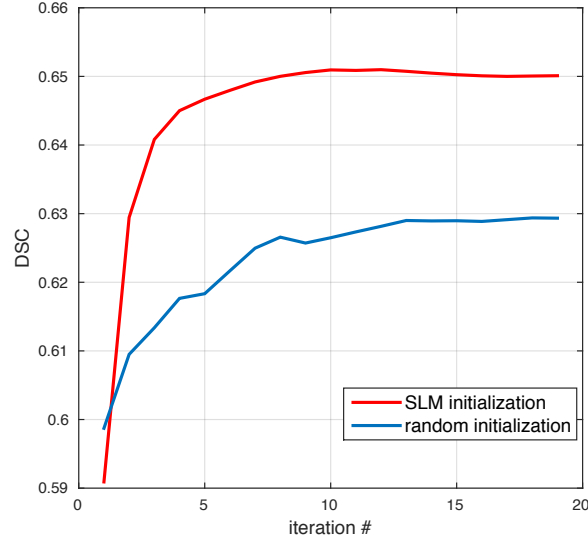


Figure 4.3: SSLM’s performance influenced by different initialisation strategies over the validation set.

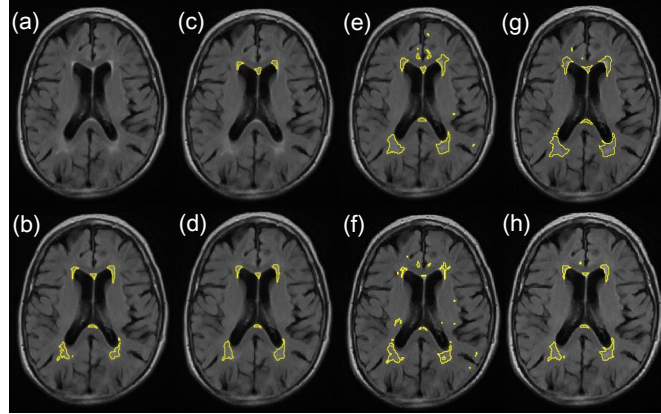


Figure 4.4: Example WMH segmentation results on subject with medium lesion load compared with LGA, LPA, UOCL, SVM and human expert annotations. (a) Original FLAIR image (b) Expert segmentation (c) LGA (d) LPA (e) UOCL (f) SVM (g) SLM (h) CS³LM

4.5.2 Experimental results on WMH segmentation

We first show a visualisation of the proposed method compared with LGA, LPA, UOCL and SVM in Fig. 4.4 and Fig. 4.5 on two different subjects with different lesion loads. Qualitative comparisons of these methods show that the proposed SLM and CS³LM framework gave a visually more accurate segmentation and improved detection of WMH compared with the other methods, not only picking up large and contiguous regions, but also detecting small and irregular

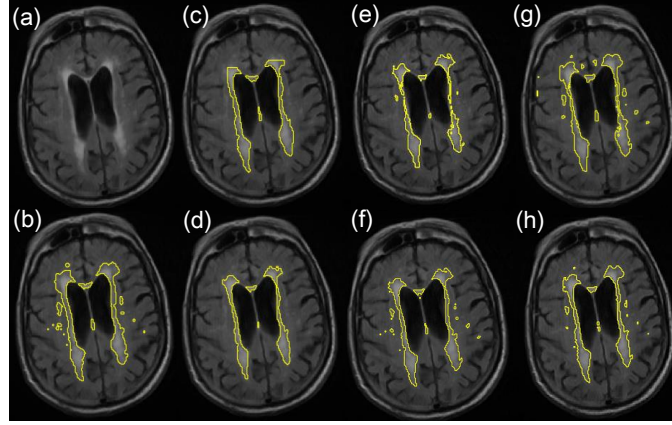


Figure 4.5: Example WMH segmentation results on subject with high lesion load compared with LGA, LPA, UOCL, SVM and human expert annotations. (a) Original FLAIR image (b) Expert segmentation (c) LGA (d) LPA (e) UOCL (f) SVM (g) SLM (h) CS³LM

Table 4.2: Performance comparison results on brain MR scans with WMH (The best results are shown in bold).

Method	DSC	ACC	Recall	Precision	Gmean
LGA	0.4259	0.9985	0.3559	0.7031	0.5557
UOCL	0.4759	0.9972	0.6802	0.4876	0.8151
LPA(0.5)	0.5106	0.9988	0.4016	0.8924	0.5995
LPA(0.2)	0.6434	0.9990	0.6447	0.7520	0.7871
SVM	0.6465	0.9990	0.6723	0.6876	0.8133
SLM	0.6505	0.9990	0.6784	0.6725	0.8166
CS ³ LM	0.6669	0.9991	0.6925	0.6860	0.8260

ones. For quantitative evaluation, a comparison of classification and overlap measures between automated segmentation results and the expertly annotated WMH masks was performed, with the results shown in Table 4.2. Here LPA(0.5) refers to thresholding the lesion map with 0.5 which is suggested by the author, while LPA(0.2) means setting the threshold to 0.2 which is chosen by searching for the best results LPA can achieve. In this experiment it can be observed that the proposed algorithm is competitive to the existing methods, and outperforms them on most of the evaluation measures. In Table 4.2, accuracy is reported for completeness sake of classification measures, however due to the nature of the problem addressed here, it offers very little insight about performance. Accuracy is strongly influenced by the number of TNs, which is inherently very large as the lesion to healthy tissue ratio is very small. The definition of recall and precision in Table 3.1 suggests that higher recall indicates more TPs, while lower precision

indicates more FPs given the same number of TPs. This shows that in comparison with LPA, the proposed method is able to produce relatively more TPs, but more FPs as well. Though LPA is good at not making FP predictions, it is more likely to miss small lesions, which might be a more difficult problem to address. In addition, DSC equally weight the number of FPs and FNs without accounting for the absolute number of TNs. Therefore, such measure is more suitable to evaluate the overall quality of lesion segmentation algorithms. Some authors regard DSC values over 0.7 as “excellent” [12], while others regard DSC values over 0.4 as “moderate”, over 0.6 as “substantial”, and over 0.8 as “almost perfect” [145]. According to those rules, the DSC analysis (Table 4.2) indicates that our degree of agreement with human expert annotations is remarkably good given that the mean DSC values on the subjects exceeded 0.6. Furthermore, the proposed SSLM step is able to further smooth the segmentation by taking the individual information on target subject into consideration while maintaining the confident prediction from a global classifier. The p -value of DSC between SLM and CS³LM is 1.54×10^{-5} . Higher DSC scores and visually more accurate results of CS³LM compared to SLM indicate the added value of the refinement steps.

With respect to the volumetric difference and correlation analysis, we normalised WMH volumes by converting them to percentage of intra-cranial volume (ICV %) in order to remove any bias introduced by differing head sizes. Regression analysis between manual segmentation volumes and automatic segmentation volumes (R^2 and Trend) and correlation analysis between segmentation volumes with visual rating scales Fazekas scores (CC Fazekas) are shown in Table 4.3. The proposed methods achieve higher determination coefficient R^2 values compared with other methods, indicating that the obtained results of SLM share around 96% variability with the expert annotations. Though R^2 value of CS³LM drops slightly, it achieves a linear trend that is closest to the ideal one. Correlation analysis with clinical variable Fazekas score also demonstrates the competitiveness of the proposed methods. However, correlation studies the relationship between two quantitative methods of measurement, not the differences, and a high correlation does not automatically imply that there is good agreement between the two methods [90]. Bland and Altman analysis (Fig. 4.6) was performed to evaluate a bias between the mean differences, and to estimate an agreement interval, within which 95% of the differences of the

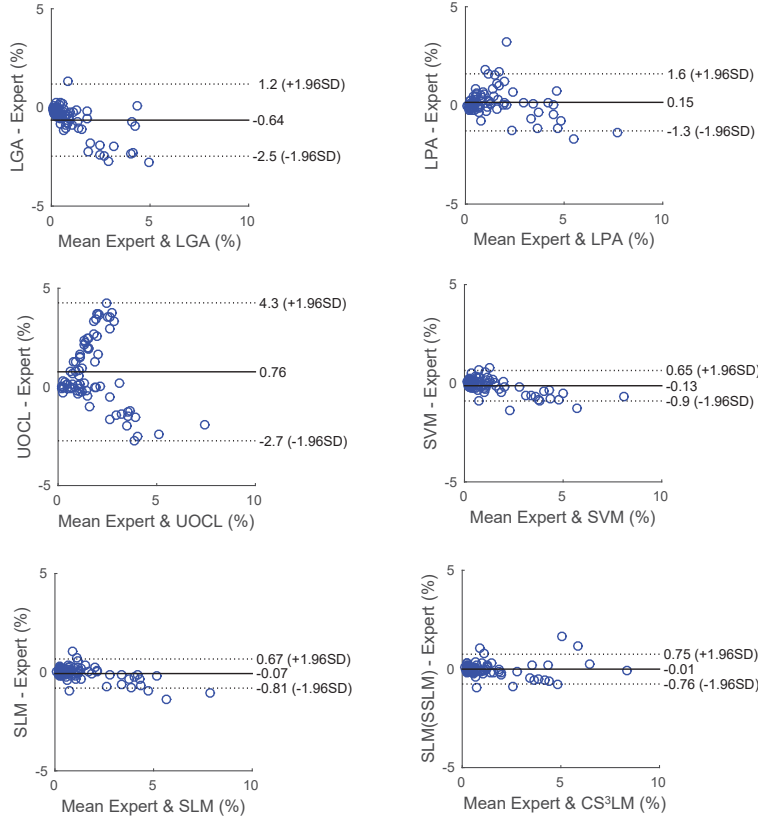


Figure 4.6: Bland and Altman plot: differences between expert annotated segmentation volumes and automatic segmentation volumes vs. the mean of the two measurements, with mean difference (solid line) and the representation of the limits of agreement (dotted line), from $-1.96SD$ to $+1.96SD$.

automatic segmentation fall compared to expert annotation [90]. Results of Bland and Altman analysis [35] are shown in Fig. 4.6, in which mean difference (solid line) and the representation of limits of agreement (dotted line), from $-1.96SD$ to $+1.96SD$ (SD: standard deviation) are given. We observed that CS³LM shows comparable and competitive results with respect to the measurement, with small mean difference and limit intervals, suggesting its high consistency with expert annotations.

4.6 Conclusions

In this chapter, novel supervised and semi-supervised large margin algorithms for WMH segmentation on MR scans are introduced. The proposed model can discover WMH first via a

Table 4.3: Correlation analysis with clinical variables (CC Fazekas) and correlation analysis between expert and automatic volumes (R^2 and Trend) (The best results are shown in bold)

Methods	CC Fazekas	R^2	Trend
Expert	0.8567	-	-
LGA	0.7782	0.7790	$0.53x+0.02$
LPA(0.2)	0.8018	0.8116	$0.78x+0.46$
UOCL	0.1065	0.1201	$0.28x+1.77$
SVM	0.8414	0.9630	$0.83x+0.11$
SLM	0.8771	0.9641	$0.85x+0.14$
CS ³ LM	0.8582	0.9493	$0.98x+0.02$

global large margin classifier learnt across a group of training data, and then followed by being fine-tuned using the proposed semi-supervised method on target data. The performance was evaluated by means of comparison with segmented WMH masks derived from expert annotations in terms of overlap and, volumetric agreement and difference. Encouraging experimental results were obtained with both the qualitative visualisation results and the quantitative scores, showing the effectiveness and competitiveness of the proposed model against other existing methods.

Chapter 5

CNNs for differential segmentation of WMHs and stroke lesions

This chapter is based on:

- **C. Qin**, R. Guerrero, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M.C. Valdés-Hernández, D.A. Dickie, J. Wardlaw, D. Rueckert. White Matter Hyperintensity and Stroke Lesion Segmentation and Differentiation Using Convolutional Neural Networks. *NeuroImage Clinical*, 2018, 17: 918-934.

This is a joint work with R. Guerrero, and both authors contributed equally.

5.1 Introduction

The accurate assessment of WMH burden is of crucial importance for epidemiological studies to determine associations between WMHs, cognitive and clinical data; their causes, and the effects of new treatments in randomised trials. The problem of WMH delineation and quantification is further complicated by the fact that other pathological features (i.e. stroke lesions) often also appear as hyperintense regions. Recently, several automated methods aiming to tackle

the challenges of WMH segmentation have been proposed. Most of these methods have been specifically developed to segment WMHs in MRI but cannot differentiate between WMHs and strokes. Other methods, capable of distinguishing between different pathologies in brain MRI, are not designed with simultaneous WMH and stroke segmentation in mind. Therefore, a task specific, reliable, fully automated method that can segment and differentiate between these two pathological manifestations on MRI has not yet been fully identified.

In this chapter, a convolutional neural network (CNN) that is able to segment hyperintensities and differentiate between WMHs and stroke lesions is proposed. Specifically, we aim to distinguish between WMH pathologies from those caused by stroke lesions due to either cortical, large or small subcortical infarcts. The proposed fully convolutional CNN architecture, called uResNet, is comprised of an analysis path, that gradually learns low and high level features, followed by a synthesis path, that gradually combines and up-samples the low and high level features into a class likelihood semantic segmentation. The choice of the uResNet architecture adopts from the U-net architecture as proposed in [206], which has been shown effective for various segmentation tasks. Slightly different from [206], convolutions in the U-net architecture are replaced with residual elements [107] and the concatenations used in skip connections are replaced with summations. This minor modification is due to the observation that residual architectures have been shown to ease gradient back-propagation flow, and hence improve optimisation convergence speed and allow for deeper network training [107]. On the other hand, the replacement of concatenations with summation operations can greatly reduce the network complexity without any performance degradation for our particular task observed from preliminary experiments. Note that the design of the architecture is not the main focus of this work, thus we chose this simple but efficient network for the following study.

An important contribution of this work deals with data sampling for training. Due to the large class imbalance present in WMH segmentation, data sampling for training requires careful consideration, an issue that has received recent research focus due to its influence on the precision of segmentation [131]. Here, to mitigate class imbalance, training is done using patches, rather than dense training on whole images. Further to this, we sample patches that always contain WMHs and randomly shift the central location so that WMHs can occur

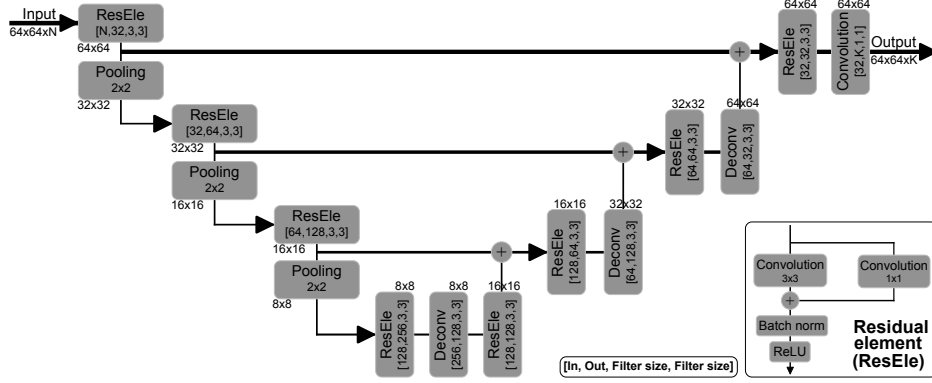


Figure 5.1: Proposed u-shaped residual network (uResNet) architecture for WMH segmentation and differentiation.

anywhere in the patch and not necessarily at the center. As argued before, the proposed CNN architecture is designed for 2D images and it is trained with 2D image patches. Furthermore, multi-channel inputs are experimented to evaluate the added benefit of adding T1 MR scans and white matter and/or cerebro-spinal track probability maps. The proposed architecture, referred as uResNet, is visualised in Fig. 5.1.

Quantitatively, the proposed CNN architecture is shown to outperform other well established and state-of-the-art algorithms in terms of overlap with manual expert annotations. Clinically, the extracted WMH volumes were found to correlate better with the Fazekas visual rating score than competing methods or the expert-annotated volumes. Additionally, a comparison of the associations found between clinical risk-factors and the WMH volumes generated by the proposed method, were found to be in line with the associations found with the expert-annotated volumes.

5.2 Methods

CNNs represent a versatile class of machine learning models that can be trained to predict voxel-wise semantic labels on images. This is achieved by learning a mapping function $f(\Theta, x) \rightarrow y$, parameterised by Θ , that transforms voxel level image intensity x to a desired label space or image segmentation $y \in Y$. Such mapping function $f(\Theta, x)$ is modelled by a series of L

convolution and non-linearity operations, with each element in this sequence generally referred to as a layer. Each layer l produces a set of feature maps H_l . Here, the convolutional kernel of layer l that produces the j th feature map is parameterised as $w_l^{j,k}$, where k refers to the k th feature map of H_{l-1} . The solution to this problem estimates a conditional distribution $p(y|x)$ that minimises the loss function Ψ (see Section 5.2.2) defined by y and its estimate $f(\Theta, x)$. After each layer l a set of feature maps or intermediate representations h_j^l is obtained. In this work, non-linearities are defined as rectified linear units (ReLU) [181]. Intermediate feature maps are computed as convolutions between the convolution kernels $w_l^{j,k}$ and the layers' input as

$$h_l^j = \max \left(0, \sum_{k=1}^{J_{l-1}} h_{(l-1)}^k * w_l^{j,k} \right). \quad (5.1)$$

Here $*$ denotes the convolution operator, $h_0^j = x$, and J_{l-1} is the number feature maps in layer $l - 1$, with J_0 being the number of input channels.

In addition to the sequence of convolution and non-linearity operations mentioned, in the work presented here, residual units or residual elements (ResEle) [107] are employed to reformulate the previous mapping function as $f(\theta_l, H_{l-1}) + W_l H_{l-1} \rightarrow H_l$, where W_l performs a linear projection that matches the number of feature maps in layer $l - 1$ to those in layer l . Figure 5.1 bottom-right shows the form of ResEle used in this work. Furthermore, to decrease the number of parameters (and control over-fitting) associated with an increased network field-of-view, max-pooling layers are employed. Max-pooling operates independently on each input feature map where all but the maximum valued activation within a support region are discarded, and the same is repeated at every strided location. Support region and stride in this work were set to 2×2 and 2 respectively, effectively down-sampling by a factor of two after every max-pool layer.

5.2.1 Network architecture

Defining a CNN's architecture requires careful consideration of the task set out to achieve. Important aspects that must be taken into account are the network's field of view or receptive

field and its capacity or complexity. In the architecture proposed here we follow the suggestions of Simonyan and Zisserman [228] and use only small (3×3) kernels. This allows an increased non-linearity capacity with a lower number of parameters needed for the same receptive field.

The architecture proposed here follows a U-shaped architecture. No fully connected layers are used, thus it is a fully convolutional network, and hence even though it is trained with image patches, inference can be performed on whole images in one single feed forward pass without any need of architectural changes. In total, our architecture is composed of 12 layers with $\sim 1\text{M}$ trainable parameters: 8 residual elements, 3 deconvolution layers, and one final convolution layer that converts the feature maps to the label space. Here, the last layer's feature maps H_L are passed to an element-wise softmax function that produces pseudo class probability maps as

$$\rho_c(H_L) = \frac{\exp(H_L)}{\sum_{c=1}^C \exp(H_L)} \quad \forall c, \quad (5.2)$$

where c denotes class and C is the total number of classes.

This in essence yields a class-likelihood for each voxel in the image, and its output, in combination with a loss function (described in Section 5.2.2), is optimised through the back-propagation algorithm.

5.2.2 Loss function and class imbalance

In general terms, a loss function maps the values of one or more variables onto a real number that represents some “cost” associated with an event. Loss functions defined for classification tasks are functions that calculate a penalty incurred for every incorrect prediction. As mentioned before, casting a semantic segmentation task as a voxel-wise classification problem tends to lead to significant class imbalances. Loss functions can be defined in such a way that they take class imbalance into account. Here, we will detail a classical loss function that does not take into account class imbalance as well as several recently proposed loss functions that either directly or indirectly take into account class imbalance.

In the context of the work presented here, let us define a training set of samples $X = \{x_1, \dots, x_P\}$, where each $x_p = \{x_{(p,1)}, \dots, x_{(p,V)}\}$ are image patches extracted from in-plane FLAIR (and/or additional modalities) axial slices that will be treated as independent samples during training. Here, P is the total number of patches available and V is the total number of voxels per patch. Additionally, let us also define voxel level labels as one-hot encoded variables $y_{p,v}$ associated with each voxel $x_{p,v} \in X$. Let us consider $Y \in \mathbb{N}^C$ a one-hot encoded label space, where the class of each voxel in $x_{p,v}$ is given by a C -length vector $y_{p,v}$ of all zeros except for a one at position c which indicates the associated label. However, let us simplify notation for the following loss equations by re-indexing all voxels in X and their corresponding label as x_n and y_n , respectively. Here, $n = \{1, \dots, N\}$ and $N = P * V$ is the total number of voxels from all patches in X . Therefore, the problem of estimating the mapping function $f(\Theta, x_n)$ can be defined as the minimisation of a loss function that works with the pseudo probabilities obtained from Eq. 5.2.

A popular loss function for classification tasks, such as the one tackled here, is the categorical cross-entropy which aims to maximise the log likelihood of the data or, equally, minimise the cross-entropy via the following loss function

$$\Psi = - \sum_{n=1}^N y_n \log(f(\Theta, x_n)). \quad (5.3)$$

Classical cross-entropy does not take into account class imbalances in the data which might lead to learning biased predictors. A simple approach to deal with class imbalance that has been proposed for CNN segmentation, is to modify the aggregation of categorical cross-entropy given in Eq. 5.3, by weighting voxels that belong to different classes differently. This modification aims to give more weight to under-represented classes, while weighting down over-represented ones, and can be written as

$$\Psi = - \sum_{n=1}^N y_n \log(f(\Theta, x_n)) \omega(y_n). \quad (5.4)$$

where $\omega(y_n)$ is the weight associated to class of y_n .

Wu et al. [272] recently proposed a simple modification of the categorical cross-entropy by dropping, or ignoring, the loss contribution of elements whose correct class prediction was above a certain threshold τ . This has the effect of placing more emphasis on previous mistakes, thus focusing the learning process on “harder” (and arguably more valuable) examples during training. Dubbed online bootstrapped categorical cross-entropy, this loss function can be written as

$$\Psi = - \sum_{n=1}^N y_n \log(\varphi_n) \quad (5.5)$$

where $\varphi_n = \{1 \text{ if } f(\Theta, x_n) > \tau, f(\Theta, x_n) \text{ otherwise}\}$.

The Dice coefficient is defined on a binary space and aims at maximising the overlap between regions of the same class. This makes it a popular and natural choice of metric when comparing binary segmentation labels. However, it is non-differentiable, making its optimisation with the back-propagation algorithm not possible. Recently, the winning team of the Second Annual Data Science Bowl¹ proposed to use a pseudo Dice coefficient as loss function, which can be written as

$$\Psi = 1 - \frac{1}{C} \sum_{c=1}^C \left(\frac{2 \sum_{n=1}^N (y_n^c f(\Theta, x_n)^c)}{\sum_{n=1}^N f(\Theta, x_n)^c + \sum_{n=1}^N y_n^c} \right). \quad (5.6)$$

Here, the predicted binary labels are replaced by continuous softmax outputs and averaged across all labels C , and $f(\Theta, x_n)^c$ denotes the softmax prediction of class c . Aggregating Dice coefficients from C different classes as an average, has the additional effect of normalising the per-class loss contribution.

5.2.3 Data sampling and class imbalance

Generally, in the segmentation of pathologies, healthy tissue is present in far larger quantities than pathological tissue. For example, in WMH segmentation the number of voxels labelled as WMH (regardless of the underlying pathology) is very small compared to those labelled background/healthy tissue, which leads to a significant class imbalance ($\sim 99.8\%$ of the voxels

¹<https://www.kaggle.com/c/second-annual-data-science-bowl>

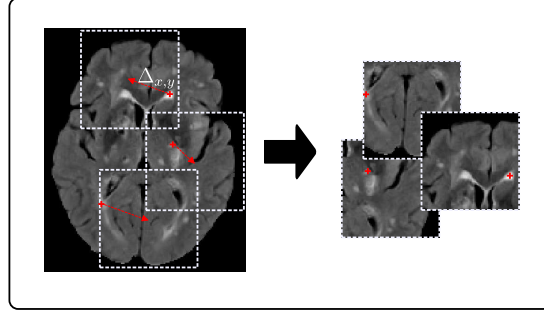


Figure 5.2: Training patch sampling.

in the dataset used in this work are labelled as background/healthy tissue in our training set). Hence, although dense training (where whole images or slices are used) is a staple in computer vision with natural images [160], it is less intuitive for WMH segmentation. Therefore, patch sampling is used in this work in order to alleviate the class imbalance problem. There are several techniques that could be used to sample patches for training. For example half of the samples could be extracted from locations centered on healthy tissue and half centered on WMH tissue [131], however this strategy does little for the class imbalance when large patches are being considered, as individual patches tend to still be highly class imbalanced at voxel level. Another option, is to sample patches centered at WMH locations only, which in fact reduces the healthy tissue class to $\sim 90\%$. However, both strategies, in combination with the proposed architecture that has a field of view comparable to sample size, would lead to a location bias, where WMHs are always expected in the center of a patch. Instead, we propose that after defining a random subset of WMH voxels from which to extract training patches, a random shift $\Delta_{x,y}$ of up to half the patch size can be applied in the axial plane before patch extraction to augment the dataset. Fig. 5.2 details this procedure. It is important to point out that the location sensitivity mentioned here, is generally not an issue with dense training in natural images, where different classes can either appear anywhere in a scene (e.g. a face might be located anywhere), or class location gives a meaningful description (e.g. sky tends to be in the upper part of a scene). This problem only occurs when sampling patches from training images in a systematic way, such as proposed here.

5.3 Data

The proposed methodology was evaluated using a subset of 167 images from 250 consecutive patients who presented themselves to a hospital stroke service with their first clinically evident non-disabling lacunar or mild cortical ischemic stroke [247]. Diabetes, hypertension, and other vascular risk factors were not criteria for exclusion. However, patients with unstable hypertension or diabetes, other neurological disorders, major medical conditions including renal failure, contraindications to MRI, unable to give consent, those who had hemorrhagic stroke, or those whose symptoms were resolved within 24 hours (i.e., transient ischemic attack) were excluded. The subset of 167 subjects considered in this work consisted of those for which all WMHs and stroke lesions were delineated (see Section 5.3.1) as different annotation classes, i.e. those that contained strokes but were not labelled as such were excluded. In this work, stroke lesions included both old and recent lesions as defined in [247], which in turn are either of cortical or sub-cortical nature.

A subset of 126 from the 167 subjects used, contained additional complete clinical and demographic data. Information included risk factors and clinical assessments such as: age, sex, reported diabetes, reported hypertension, reported hyperlipidaemia, reported smoking, mini mental state examination (MMSE), systolic blood pressure (SBP), diastolic blood pressure (DBP), total cholesterol, peri-ventricular Fazekas score (PV-Fazekas), deep white matter Fazekas score (D-Fazekas), deep atrophy volume (deepAtrophyVol), basal ganglia enlarged peri-vascular spaces (BGPVS) score, centrum semiovale enlarged peri-vascular spaces (CSPVS) score, old stroke lesion (oldLes) present, and total number of micro-bleeds (micrBld).

5.3.1 MRI acquisition

All image data was acquired at the Brain Research Imaging Centre of Edinburgh ² on a GE Signa Horizon HDx 1.5T clinical scanner (General Electric, Milwaukee, WI), equipped with a self-shielding gradient set and manufacturer-supplied eight-channel phased-array head coil.

²<http://www.bric.ed.ac.uk>

Details of the protocols used for acquiring the data are given in Table 5.1, and their rationale is explained in [247]. Although several imaging sequences were acquired, only T1 and FLAIR MR images were used for this study. Of the 167 subjects considered in this work 35 were acquired under protocol 1, 83 under protocol 2 and 49 under protocol 3.

Protocols	Protocol 1	Protocol 2	Protocol3
TR/TE/TI (ms) T1	9/440	9.7/3.984/500	
TR/TE/TI (ms) FLAIR	9002/147/2200	9000/140/2200	
Pixel bandwidth (KHz)	125 (T1) 122.07(FLAIR)	15.63 (T1) 15.63 (FLAIR)	
Matrix	256 × 192	256 × 216 (T1) 384 × 224 (FLAIR)	192 × 192 (T1) 256 × 256(FLAIR)
No. slices	20	256 (T1) 28 (FLAIR)	160 (T1) 40 (FLAIR)
Slice thickness (mm)	5	1.02 (T1) 5 (FLAIR)	1.3 (T1) 4 (FLAIR)
Inter-slice gap (mm)	1.5	1	0
Voxel size (mm ³)	0.94 × 0.94 × 6.5	1.02 × 0.9 × 1.02 (T1) 0.47 × 0.47 × 6 (FLAIR)	1.3 × 1.3 × 1(T1) 1 × 1 × 4 (FLAIR)

Table 5.1: MR imaging sequence details for the three acquisition protocols used.

Image pre-processing and gold standard annotations

All image sequences (from each patient) were co-registered using FSL-FLIRT [125] and mapped to the patient’s FLAIR space. Hyperintensities from MR images that were acquired under protocol 2 (Table 5.1) were delineated using Multispectral Coloring Modulation and Variance Identification (MCMxxxVI). Described in [247, 248], MCMxxxVI is based on the principle of modulating or mapping two or more different MRI sequences that display the tissues/lesions of the brain in different intensity levels. The implementation of the technique can be found in www.sourceforge.net/projects/bric1936. It fused different co-registered MRI sequences and then applied minimum variance quantisation as the clustering technique to segment different tissue types. Here, MCMxxxVI considered hyperintensities as those hyperintense signals that simultaneously appear in all T2-weighted based sequences. Different from protocol 2, hyperintensities from MR images acquired under the protocols 1 and 3 were delineated via histogram-based thresholding of the FLAIR sequences followed by human correction. These hyperintense regions were further separately classed as WMHs, old stroke lesions, and recent

stroke lesions using the Region of Interest tool in Analyze 11.0TM. The tissue loss due to stroke and cavities is separately delineated semi-automatically by combining thresholding and an interactive region-growing algorithm, while guided by radiological knowledge, on FLAIR (if ischemic) or T2-weighted (if hemorrhagic) images [247, 248]. Their identification procedure is described in the Table 2 of [247] and single stroke class was created by combining recent and old. All images were re-sliced as to have $1mm$ in both dimensions of axial slices, with the remaining dimension (slice thickness) left unchanged. White matter probability maps were obtained from T1 image segmentation using [150] and cerebro-spinal track probability maps were obtained by co-registering a tract probability map [115] to the FLAIR image space. Additionally, in order to have consistent intensity voxel values for model training, all MR images were normalised to have zero mean and standard deviation of one (excluding the background). Values below three standard deviations from the mean were clipped in order to guarantee consistent background values across all images.

5.4 Experiments and results

Data used as input for training the proposed CNN (uResNet) was sampled from whole brain MR volumes as explained in Section 5.2.3. Image patches and the corresponding labels of 64×64 voxels were extracted from the volumes at a random subset of 20% of all possible locations that were labeled as WMHs and 80% of locations labeled as stroke. All 167 images included in this study contained WMH lesions, of these, 59 also contained stroke lesions. Data was split into two separate sets used for two fold cross-validation, where each fold contained half of the images with WMHs only and half with both WMHs and stroke, as to represent data distribution in both folds. During each fold of the cross validation experiments, one fold is used for training (network parameter learning) and setting all other parameters, while the second (unseen) fold is reserved for testing. That is, optimisation of the loss function, input channel selection and stopping criteria are carried out on the training set. Section 5.4.3 shows a comparison of the proposed uResNet with a version that used residual blocks with two convolutions (uResNet2) and to observe the added value of the center shifting during training patch sampling (uResNet_NoC).

Experiments were carried out using the Theano [11] and Lasagne [68] frameworks with Adam [138] optimisation (default Lasagne parameters), mini-batch size of 128, learning rate of 0.0005 (with 0.1 decay after 25 epochs) and random weight initialisation (default Lasagne settings).

The evaluation criteria used to compare all methods can be split in two, mainly, a comparison to other well established and state-of-the-art methods and a clinical analysis. The comparison to other methods consisted of an evaluation of the labelling overlap of the segmented images using the Dice coefficient, and an analysis of the differences between the automatically calculated volumes and the expert in terms of intra-cranial volume (ICV). Comparison results calculated using the Dice coefficient and volume analysis are reported on a per class basis. Clinical evaluations consisted of a correlation analysis with some clinically relevant variables (mainly Fazekas and MMSE scores), and a general linear model (GLM) analysis of association with known risk factors.

5.4.1 Model training

An important factor during CNN training is the definition of the loss function that will guide the learning process (Section 5.2.2). Here, we experimented with several recently proposed loss functions that were used to train the proposed WMH segmentation CNN network using FLAIR images as input. In order to directly compare the effect of different loss functions, Dice score results from evaluating the CNN after different stages of training were calculated, see Fig. 5.3. Here, the horizontal axis indicates the number of training epochs while the vertical axis indicates the Dice score achieved on either the train (top row) or test (bottom row) datasets. In this work, an epoch is defined as transversing the complete set of training of patches once. It must also be noted that Dice results displayed here are calculated on the whole brain MR volumes, not on the extracted patches. Fig. 5.3 shows the results obtained using classical, bootstrapped and weighted cross-entropy loss functions, as well as using a pseudo Dice similarity score (see Section 5.2.2). From the top row of Fig. 5.3 (results on train data) it can be observed that weighted and classical cross-entropy perform best and that there is little difference between them. However, weighted cross-entropy has an additional class weight

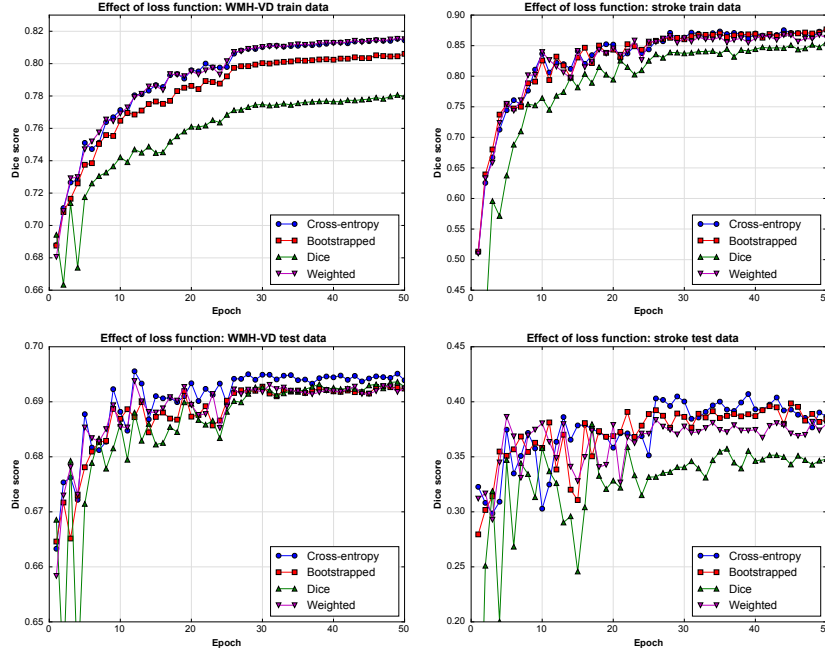


Figure 5.3: Effect of different loss functions on uResNet trained using FLAIR images.

parameter associated that needs to be set. Hence, for the problem presented in this work and considering the experiments conducted, classical cross-entropy was considered the best choice. It is important to take notice that using the Dice coefficient as both loss function and evaluation metric provides surprisingly poor results during training (top row Fig. 5.3). Here, we theorise that, for this particular problem, the solution space over which we optimise might be more complex for the Dice metric than the other, and hence finding a global optimal solution might prove more cumbersome.

As mentioned before, WMHs are best observed in FLAIR MR images, however it has been suggested that complementary information might be found on T1 MR images. In this work, the contribution from additional multi-modal information to the proposed segmentation framework was explored. Additional input channels to the proposed CNN include T1 MR images, white matter probability maps and a cerebro-spinal tract atlas. Segmentation accuracy is again evaluated using the Dice score. From Fig. 5.4, it can be seen that training converges after about 30 epochs, that is, traversing the whole set of extracted training patches 30 times. Therefore, test Dice scores and automatic volumes further presented here are obtained evaluating the model at 30 epochs.

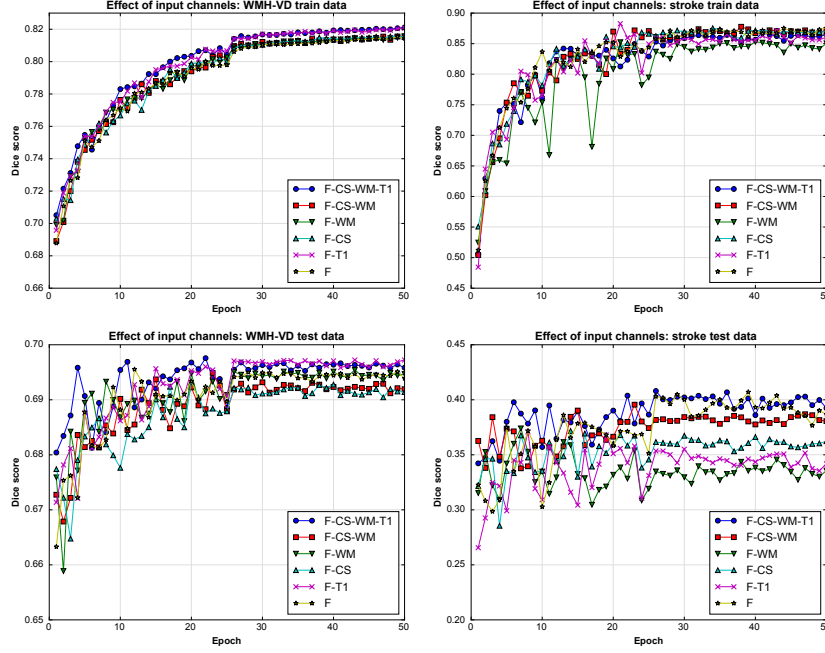


Figure 5.4: Different input channel exploration. F: FLAIR image, CS: cerebro-spinal track atlas, WM: white matter probability map, T1: T1 weighted image.

Given the different input channels, training data and testing results that take into account both segmentation classes (shown in Fig. 5.4) indicate that there is very little difference between using four input channels (FLAIR, T1, WM and CS) compared to just using FLAIR images. Hence, all subsequent experiments made use of only FLAIR images as input channels. This is additionally justified by the fact that some of the comparison methods only use FLAIR images. Furthermore, the acquisition of additional modalities (T1) or probability map generation (WM and CS) can be costly/time consuming and render the methodology less clinically viable. In Fig. 5.3 and Fig. 5.4 it can also be observed that training and testing Dice scores for stroke segmentations are much more oscillatory than those from WMH segmentation. This behavior can be explained by the fact that there is simply a lot less data of the stroke class, in fact there are ~ 14 times more WMH voxels. Therefore, stroke results are more sensitive to variations in the network's parameters as each epoch provides more stochastic gradients associated to this class. Furthermore, on the stroke dataset, higher training accuracy combined with the lower test accuracy can be attributed to this class imbalance as they potentially point to an over-fitting problem.

5.4.2 Comparison to state-of-the-art

In the experiments presented in this section, the proposed uResNet was compared to other well established and state-of-the-art algorithms. From the lesion segmentation toolbox (LST) version 2.0.15³, the LPA and LGA frameworks were used. LPA used only FLAIR images as input while LGA required both FLAIR and T1 images. DeepMedic, a recently published CNN library for segmentation of medical images, was also used in the comparisons presented here with its default settings. Parameters for both LPA and LGA frameworks were set according to a two fold cross-validation using the same data splits as described before for uResNet. LPA has only one parameter, a threshold τ used to binarise the lesion probability maps generated, and the optimal value τ after cross-validation was set to 0.16. The authors recommend setting this value to 0.5, however this produced poor results and hence were excluded from further analysis. LGA has two parameters, κ that was set to 0.12 after cross-validation and a threshold that was set to 0.5. DeepMedic was also validated using the same two fold cross-validation strategy (with FLAIR images as input), where the network is trained in one fold and tested in the other, however, no other meta-parameter (e.g. network’s filter sizes, number of feature maps or learning rate) tuning was done. DeepMedic was trained using images re-sliced to isotropic $1mm^3$ voxel size, and patch sampling was internally handled by DeepMedic. The default sampling option was used, which randomly samples 50% of patches from healthy tissue and 50% from pathological tissue (without considering different class weightings).

Dice overlap scores between automatically generated and expertly annotated WMHs and stroke lesions are shown in Table 5.4. Here, it can be observed that the proposed uResNet outperforms the compared methods, with all comparisons between the Dice scores obtained with the proposed and every competing being found to be statistically significant $p < 0.01$ according to Wilcoxon’s signed rank test. Statistical significance gives a measure of the likelihood that the difference between two groups could be attributed to chance, while effect size (or the “strength of association”) quantifies the relative magnitude of the difference between those two groups. Cohen [55] describes effect size values of 0.2, 0.5 and 0.8 as small, medium and

³<http://www.statistical-modelling.de/lst.html>

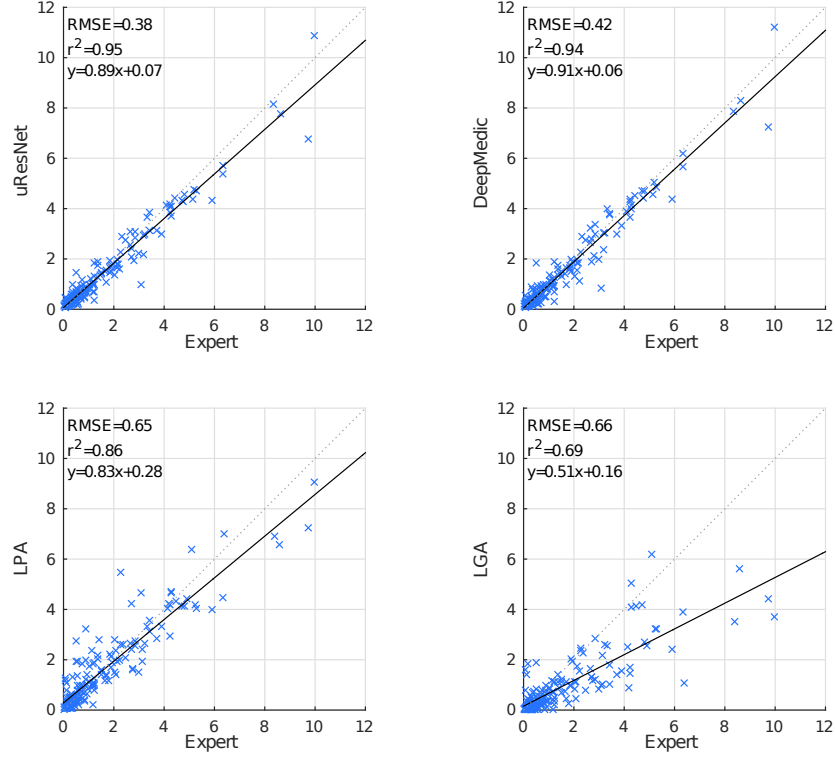


Figure 5.5: Automated versus expertly generated WMH volumes (as ICV %). The solid line indicates the linear trend $f(x)$ of the comparison, while the dotted line indicates the ideal trend $f(x) = 1.0x + 0.0$.

large, respectively. Effect sizes related to the statistical significance tests were calculated, and were 0.45, 0.32 and 0.61 for the comparison of uResNet Dice scores against those from DeepMedic, LPA and LGA, respectively. Fig. 5.5 shows a correlation analysis between the expertly annotated WMH volumes and those automatically generated. To remove any potential bias associated with head size and thus allow a better comparison, volumes were converted to ICV %. Ideally, automatic algorithms should produce values as similar as possible to the expert, and hence, should lie close to the dotted lines in Fig. 5.5. The solid lines indicate the general linear trend of the expert vs. automatic comparison and the coefficient of determination R^2 indicates to what degree automatic values explain the expert ones. From Fig. 5.5 bottom row we can see that both LPA and LGA perform clearly worse than the CNN approaches (uResNet and DeepMedic). It is also evident that LPA outperforms LGA, where each has a R^2 value of 0.86 and 0.69, respectively. Differences between uResNet and DeepMedic (top row of Fig. 5.5) are less evident. However, on close inspection of the R^2 metric in Table 5.4 of uResNet and

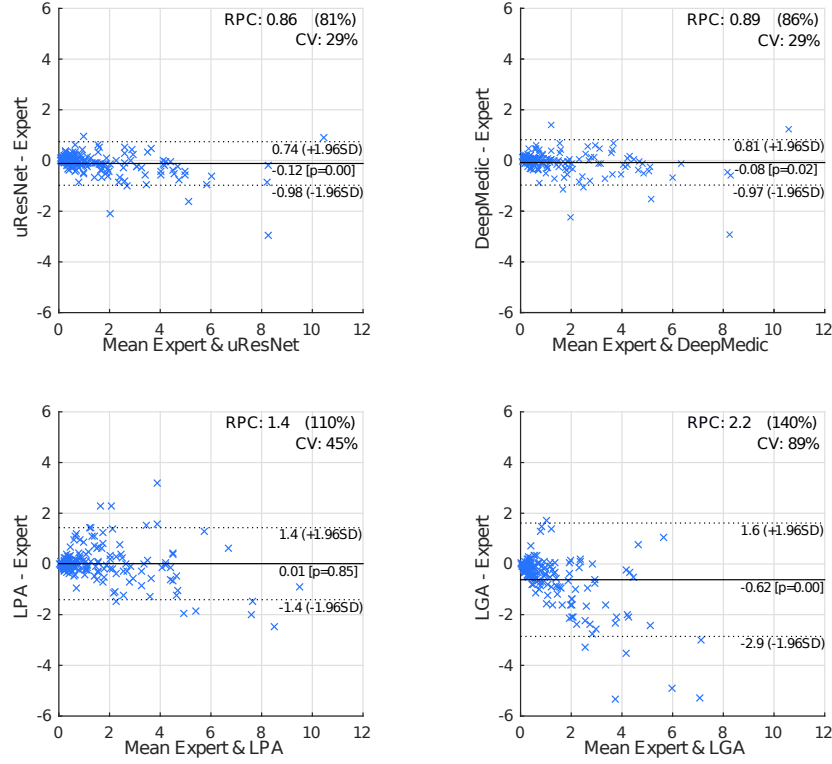


Figure 5.6: Bland-Altman plots comparing expert annotations with all other methods in WMH segmentation.

DeepMedic we can see that uResNet results are slightly better correlated to those generated by the expert. On the other hand, DeepMedic has a slope of 0.91 (offset 0.06) while uResNet has a slope of 0.89 (offset 0.07), suggesting a slightly better agreement.

Fig. 5.6 shows Bland-Altman plots that further compare expert and automatic WMH volumes. In these plots, the horizontal axis gives the average between expert and automatic volumes for each subject, while the vertical axis shows the difference between these volumes. The reproducibility coefficient (RPC), as calculated here, gives a measure of the variability (or spread) of the differences between automatic and manual volumes and is calculated as 1.96 times the standard deviation σ of those differences ($1.96 * \sigma$). In the experiments presented here, smaller values indicate better agreement between automatic and manual volumes. The coefficient of variation (CV) is given by $100 * \sigma / \bar{X}$, where $\bar{X}$ refers to the mean volume from both measurements. Dotted lines in the plots of Fig. 5.6 give the range of the RPC. Bland-Altman plots also provide insight into possible biases of compared methods. LGA displays a

statistically significant tendency to under-estimate volumes (central solid line). However, all methods tend to under-estimate larger volumes and over-estimate small ones, with the effect more pronounced in LGA.

One of the main objectives of the work presented here is to differentiate between WMHs and stroke lesions. Neither LPA or LGA are capable of making such a distinction, and therefore are not suitable algorithms for this problem. Fig. 5.7 (top-row) shows the correlation analysis between automatic (uResNet and DeepMedic) and expert stroke volumes (normalised as ICV). It is evident that uResNet outperforms DeepMedic in terms of RMSE, R^2 and linear fit slope. Further to this analysis, Fig. 5.7 (bottom-row) shows Bland-Altman plots that further confirm these findings, where uResNet obtains a smaller RPC and CV than DeepMedic, with neither method on average displaying a significant tendency to over- or under-estimate volumes (see central solid line on plots). However, it is worth noting that both methods have a tendency to over-estimate small volumes and under-estimate larger ones. A summary of Fig. 5.5 and Fig. 5.7 is also presented in Table 5.4, where a difference between both algorithms in terms of Dice scores can be observed. Statistical significance between the comparison of uResNet and DeepMedic Dice scores was found to be $p < 0.05$ according to Wilcoxon's signed rank. with an effect size related to this statistical significance (as suggested by [186]) of 0.12. The gap between uResNet and DeepMedic can be considerably closed if additional inputs are provided to DeepMedic, as shown in Table 5.2, where different input channels are provided for both methods. It can be appreciated that DeepMedic can narrow the Dice overlap gap with uResNet if several inputs are provided. However, obtaining and generating these extra inputs limits clinical applicability and also adds additional computational costs to the whole segmentation framework, as this requires an additional MR image acquisition (and co-registration of such image), tissue segmentation and/or co-registration of a cerebro-spinal track atlas.

Fig. 5.8 shows the segmentation results from three example subjects that illustrate the differences between the methods. Here, it can be observed that uResNet generally does a better job at differentiating between WMHs and stroke lesions when compared to DeepMedic (top and middle row). In the bottom row of Fig. 5.8, an example is illustrated when uResNet wrongly segments some WMHs as stroke. Additionally, in the top row, all methods are shown to clearly

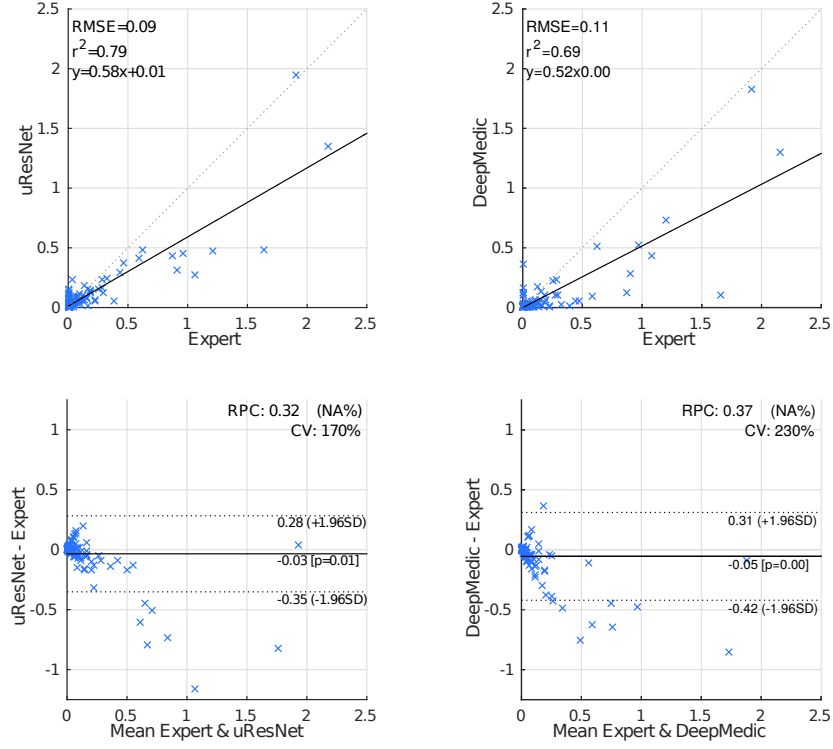


Figure 5.7: Automated versus expertly generated stroke volumes. LPA and LGA are unable to distinguish between WMH and stroke, hence cannot generate these results. The solid line indicates the linear trend $f(x)$ of the comparison, while the dotted indicates the ideal trend $f(x) = 1.0x + 0.0$.

Input channels	WMH			Stroke		
	uResNet	DeepMedic	Diff	uResNet	DeepMedic	Diff
F	69.5	66.3	<i>3.2</i>	40.0	31.1	<i>8.9</i>
F-T1	69.7	67.6	<i>2.1</i>	35.5	34.3	<i>1.2</i>
F-CS	69.1	66.6	<i>2.5</i>	36.7	35.1	<i>1.6</i>
F-WM	69.4	68.2	<i>1.2</i>	33.0	35.9	<i>-2.9</i>
F-CS-WM	69.3	68.0	<i>1.3</i>	38.4	37.8	<i>0.6</i>
F-T1-CS-WM	69.6	68.4	<i>1.2</i>	40.2	36.0	<i>4.2</i>

Table 5.2: Mean Dice scores of WMH and stroke, for different inputs with uResNet and DeepMedic. Difference in Dice score between the two methods is given in italics. F: FLAIR image, CS: cerebro-spinal track atlas, WM: white matter probability map, T1: T1 weighted image.

under-segment the image when compared to the expert. However, inspecting the FLAIR image of this subject (top row, leftmost column), it can be seen that the under-segmented regions would be challenging even for another expert annotator.

5.4.3 Variations of uResNet

In this section we present results comparing the proposed architecture and sampling scheme, with two additional versions: One where the residual block takes the more traditional form of two convolutional elements (called uResNet2) and another where the proposed center shifting sampling scheme is replaced with a standard centered patch sampling scheme (called uResNet_NoC, for not off-centered). Table 5.3 summarises these results.

Using single convolution residual blocks was noted by He et al. [107] to be equivalent to a linear projection. After experimented with residuals blocks of one and two convolutions, we observed no statistical difference ($p > 0.05$) between them. However, learning the residual of these linear projections might still be simpler, thus leading to an observed faster convergence. This observations need to be interpreted with care. The lack of observed difference in performance between one and two convolutions in residual blocks, might come down to limitations of the data itself.

Training with patches that always contain a diseased label in the center would bias towards labeling this region of a patch as diseased during inference. Patch center shifting alleviates this problem due to the distribution of probability to observe a lesion across the whole field-of-view. Allowing patches to be shifted spreads this probability to all locations and not any single location has a preferential likelihood of being a lesion. In a fully convolutional neural network, predictions can be made over a large area (as the network proposed here), taking into account context information from large areas of an image (the field-of-view or receptive field). However, training is driven by pixel-wise prediction errors, hence labeling occurs on a per-pixel basis. The likelihood of observing a lesion at any particular location is in fact very low and more or less uniform. It is this uniformity that removes the bias towards any particular location. Results comparing a uResNet without center shifting sampling are shown in Table 5.3. It can be seen that when removing the off-center sampling strategy, the regular no off-center sampling strategy (uResNet_NoC) has a significant performance drop compared with uResNet, and achieves similar performance with DeepMedic using its default sampling strategy. This indicates the effectiveness of the proposed sampling strategy for the dense lesion segmentation

	uResNet	uResNet2	uResNet_NoC	Expert
WMH Dice (std)	69.5(16.1)	69.6(16.1)	66.9(18.1)	-
Stroke Dice (std)	40.0(25.2)	40.2(27.7)	28.9(22.3)	-
WMH R^2	0.951	0.951	0.948	-
Stroke R^2	0.791	0.761	0.710	-
WMH Trend	0.89x+0.07	0.89x-0.08	0.89x+0.15	-
Stroke Trend	0.58x+0.01	0.55x- 0.01	0.52x+0.07	-
CC D-Fazekas	0.770	0.776	0.771	0.774
CC PV-Fazekas	0.778	0.783	0.777	0.765
CC Fazekas	0.824	0.831	0.823	0.819
CC MMSE	0.364	0.373	0.366	0.372

Table 5.3: Mean Dice scores of WMH and stroke (standard deviation in parenthesis), correlation analysis between expert and automatic volumes (R^2 and trend), and correlation with clinical variables. No statistical significance between uResNet and uResNet2 was observed ($p > 0.05$), while there was a statistically significant difference ($p < 0.001$) between patch off-center sampling (uResNet) and regular no off-center sampling (uResNet_NoC).

tasks.

5.4.4 Additional DeepMedic experiments

DeepMedic experiments that aim to approximate the sampling scheme used by uResNet were carried out, where several sampling weights were tested for DeepMedic. A direct comparison of per-class patch sampling is not straight forward between the proposed method and DeepMedic, and furthermore it can be misleading. For instance, in the work proposed here, a sampling rate of 80-20% of WHM-stroke patches is used, each patch has a size of 64×64 voxels and uResNet makes a prediction of a 64×64 patch of the label space during training (it is fully convolutional and uses padded convolutions throughout). This means that each patch used in uResNet has a label map that due to its size inevitably contains a large amount of healthy tissue. On the other hand, DeepMedic trains with segments that have a label space of $9 \times 9 \times 9$ voxels, therefore it is far less likely that healthy tissue is included in non-healthy samples and thus healthy segments need to be sampled. This also removes the need for off-centering sampling, since both healthy and non-healthy samples are sampled during the training. Nonetheless, different per-class sampling rates, as well as other hyper-parameter settings with DeepMedic were explored.

Some of DeepMedic’s default hyper-parameter values are: learning rate of $1e-3$, RmsProp optimizer, sampling form of foreground/background (diseased/healthy tissue) and sampling rate of $[0.5, 0.5]$ (healthy and diseased tissue). The different sampling rates tested with DeepMedic in our experiments to approximate uResNet setup were $[0.5, 0.1, 0.4]$, $[0.5, 0.25, 0.25]$, $[0.33, 0.13, 0.53]$ and $[0.33, 0.33, 0.3]$, for healthy, WMH and stroke tissue, respectively. Additionally, learning rate values explored were in the range of $1.9e-2$ to $1e-4$, with RMSprop, Adam or SGD as optimiser. Changing the sampling rates from the default generally produced unstable results, with either failing to converge or producing poorer overlap values than with the default settings. In total, 14 different additional DeepMedic train/test runs were performed, out of which only two converged, both using a sampling rate of $[0.33, 0.33, 0.33]$. Dice overlap results by these experiments were of 60.7 and 29.9, for WMH and stroke, respectively, in one instance and 59.4 and 29.5 in the other. These unstable results might be due to the tuning of additional meta parameters, such as the optimiser, learning rate or regularisation. Therefore, presented DeepMedic results were obtained with default hyper-parameters, which were the best results obtained in our experiments.

5.4.5 Clinical evaluation

Experiments thus far indicate a better agreement between volumes generated by uResNet and expert annotations, however, the question of the clinical validity of such results remains open. In this regard, Table 5.4 gives correlation coefficient (CC) results between the volumes and some clinical variables (Fazekas scores and MMSE). Fazekas scores were split into deep white matter (D-Fazekas) and peri-ventricular (PV-Fazekas), with values ranging from 0-3. An additional combined Fazekas score, created by adding both D-Fazekas and PV-Fazekas, is also presented. From Table 5.4 we can observe that in terms of correlation to Fazekas score the proposed uResNet outperforms the other competing methods, additionally noting that CC results for PV-Fazekas and Fazekas are even higher than those obtained from the expert annotations. However, in terms of CC with MMSE it was LPA that performed best.

Using the clinical scores as well as known risk factors available, an analysis of association

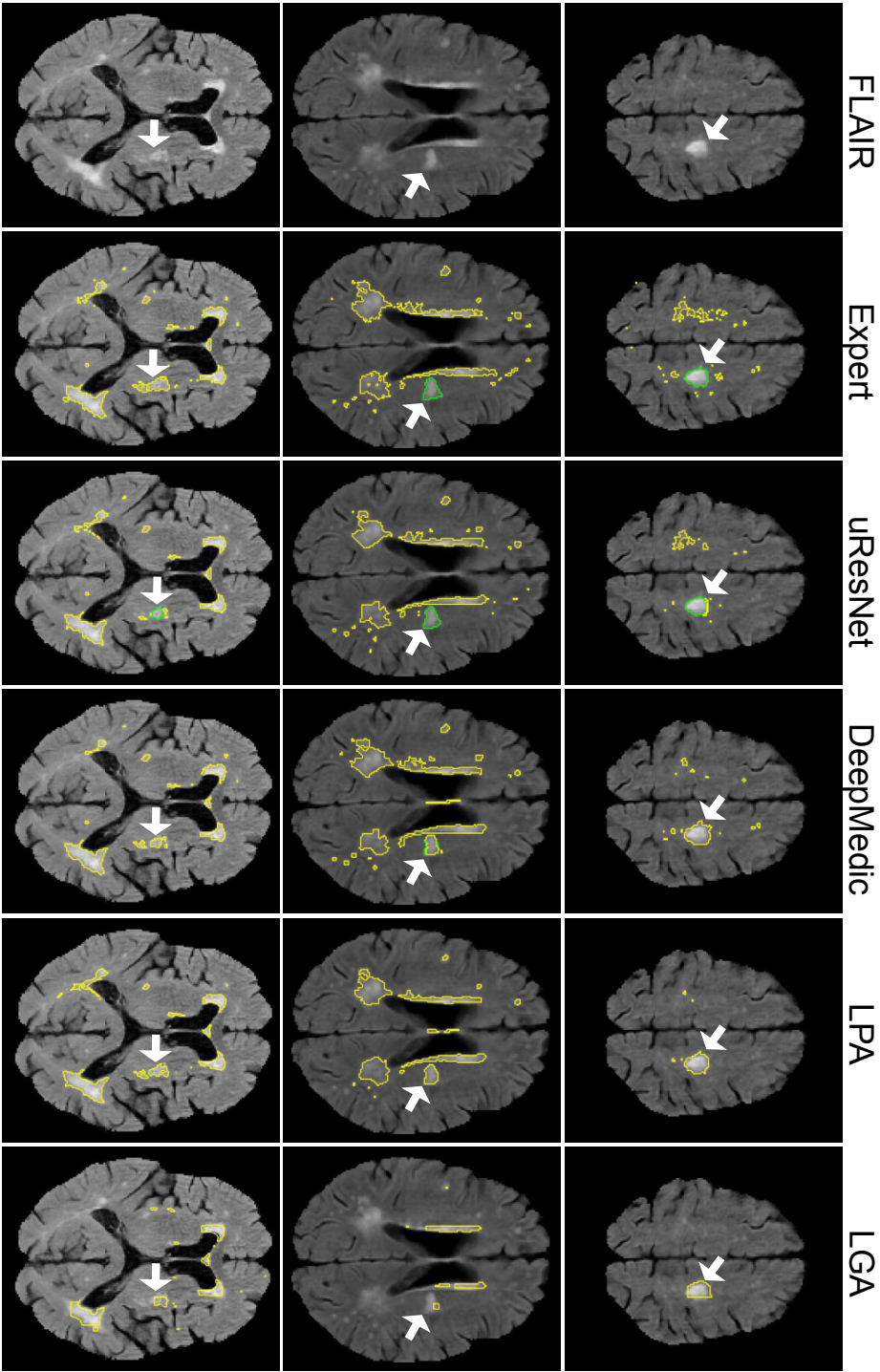


Figure 5.8: Visual comparisons of all competing methods. Yellow lines delineate WMHs, green lines stroke and white arrows point to interesting result areas. Best seen in color.

	uResNet	DeepMedic	LPA	LGA	Expert
WMH Dice (std)	69.5(16.1)	66.6(16.7)	64.7(19.0)	41.0(22.9)	-
Stroke Dice (std)	40.0(25.2)	31.3(29.2)	-	-	-
WMH R^2	0.951	0.943	0.855	0.687	-
Stroke R^2	0.791	0.688	-	-	-
WMH Trend	0.89x+0.07	0.91x-0.06	0.83x+0.28	0.51x+0.16	-
Stroke Trend	0.58x+0.01	0.52x-0.00	-	-	-
CC D-Fazekas	0.770	0.769	0.746	0.630	0.774
CC PV-Fazekas	0.778	0.780	0.777	0.718	0.765
CC Fazekas	0.824	0.824	0.811	0.734	0.819
CC MMSE	0.364	0.369	0.443	0.389	0.372

Table 5.4: Mean Dice scores of WMH and stroke (standard deviation in parenthesis), correlation analysis between expert and automatic volumes (R^2 and trend), and correlation with clinical variables.

	uResNet	DeepMedic	LPA	LGA	Expert
age	0.491	0.533	<0.001	0.723	0.313
diabetes	0.082	0.072	0.003	0.070	0.066
hyperlipidaemia	0.645	0.547	0.551	0.687	0.728
hypertension	0.820	0.781	0.504	0.358	0.562
smoking	0.497	0.560	0.216	0.719	0.767
totalChl	0.235	0.281	0.161	0.328	0.371
BGPVS	<0.001	<0.001	<0.001	<0.001	<0.001
deepAtrophyVol	0.015	0.019	<0.001	<0.001	0.117

Table 5.5: p-values of linear regression associations between volumes calculated with different methods and risk factors. Bold numbers indicate statistical significance above 0.05.

between WMH volumes and risk factors was carried out. In order to explore such associations a GLM between the results of every algorithm (as well as the expert) and the risk factors was generated. After careful consideration, age, sex, reported diabetes, reported hypertension, reported hyperlipidaemia, reported smoking, total cholesterol, deep atrophy volume and BGPVS score were used in the GLM analysis. Table 5.5 provides p-values that indicate if a particular risk factor is associated with the generated WMH volumes, where the GLMs were corrected for gender differences. Results indicate that only BGPVS is found to be associated with the expertly generated volumes, however deep atrophy volume was also found to be associated with all other methods. Additionally, LPA volumes were also found to be associated with age and diabetes.

In GLM analysis, values that are not well described by the model (outliers) can have a significant impact in subsequent analysis. Outliers in GLM can be identified by examining the

	uResNet	DeepMedic	LPA	LGA	expert
age	0.905	0.993	< 0.001	0.685	0.407
diabetes	0.012	0.019	< 0.001	0.177	0.003
hyperlipidaemia	0.346	0.425	0.464	0.550	0.186
hypertension	0.639	0.502	0.190	0.128	0.350
smoking	0.069	0.084	0.107	0.673	0.343
totalChl	0.294	0.212	0.222	0.043	0.868
BGPVS	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
deepAtrophyVol	0.005	0.008	< 0.001	< 0.001	0.020

Table 5.6: p-values of linear regression associations between volumes calculated with different methods and risk factors after residual outliers were removed. Bold numbers indicate statistical significance above 0.05.

probability distribution of the residuals. In order to eliminate any potential bias introduced by outliers, an analysis with outliers removed was performed. Results of this outlier-free association analysis are presented in Table 5.6. From Table 5.6 we can observe that once outliers were removed, expert volumes were found to be associated with deep atrophy volume, BGPVS and diabetes. The same associations were found for uResNet, DeepMedic and LPA, with the addition that LPA was again also associated with age. LGA was found to be only associated with BGPVS and deep atrophy volume.

Fazekas scores are highly co-linear with WMH volume and therefore were excluded from all previous GLM analysis. Nonetheless, a GLM that included Fazekas scores was also composed as a sanity check that the correct associations would be found. A single Fazekas score was generated by adding the D-Fazekas and PV-Fazekas scores (0-6 score scale). All models found a very strong association ($p \ll 0.001$) between Fazekas and WMH volumes. Of the expert stroke lesion volumes, systolic blood pressure was the only risk factor to be found associated ($p < 0.05$), which incidentally was also associated with the automatically (uResNet and DeepMedic) generated volumes. uResNet values were additionally found to be associated with hypertension. However, it is important to note the small size and heterogeneous nature of the population used in this analysis, which might not prove sufficient to uncover some associations. Due to the small sample analysed no outlier removal analysis was performed for stroke associations.

5.5 Discussion and Conclusions

In this work we have proposed a CNN framework, uResNet, for the segmentation of WMHs that is capable of distinguishing between WMHs arising from different pathologies, mainly WMHs of presumed VD origin and those from stroke lesions. Comparison results indicate that the proposed uResNet architecture outperforms other well established and state-of-the-art algorithms.

The architecture used in uResNet follows closely the architecture of U-Net [206]. The main difference being the use of residual elements and a generally lower complexity through the use of summation instead of concatenation in skip connections. Preliminary experiments with both summation and concatenation of features maps found no difference in performance, hence low complexity was favored. However, it is also noted that a more general solution is given by the use of concatenation, as this would allow the network to learn the best way of combining the feature maps during training. Of course this additional complexity comes at the expense of a higher risk of over-fitting and a higher memory consumption. As mentioned, the use of residual units provide advantages during training, mainly improved convergence rates in our experiments.

Large class imbalance in medical image segmentation is generally an issue that must be considered. Loss functions that take into account the class imbalance have the drawback that they have the additional class weighting parameter to tune. An additional complication resulting from a large class imbalance is that a lot of computational effort might be spent optimising to perform well in large and relatively easy to classify/segment sections of an image. Bootstrapped cross-entropy attempts to focus the learning process on hard-to-classify parts of an image by dropping out loss function contribution from voxels that have already been classified to a good degree of certainty. However, this technique also requires the setting of an additional parameter, the threshold to consider a classifications as already good, and moreover, evaluation results indicated a performance similar to classical cross-entropy.

A very important factor of the proposed CNN framework is the training data sampling

strategy described in Section 5.2.3. CNN training for medical imaging using patches is a somewhat standard technique that helps reduce the very large class imbalance that usually affects medical image segmentation. However, careful consideration must be given in the sampling strategy adopted for a certain architecture. As mentioned, class imbalance and lesion location within samples need to be considered. The use of the proposed sampling strategy described in Section 5.2.3 had a profound effect on the proposed uResNet, with WMH and stroke Dice scores increasing from ~ 67 to ~ 70 and from ~ 29 to ~ 40 , respectively, due to this alone. Another important factor is the frequency each class is sampled. In this work we sampled at 20% of the locations labeled as WMHs while at 80% of the locations labeled as strokes, again to try to balance classes. It is important to note that the default sampling settings of DeepMedic were used as in [131]. In this default sampling strategy, DeepMedic samples equally from healthy and diseased tissues (that is without considering frequency of different diseased classes) and furthermore does not include the central voxel offset sampling strategy used here. We believe that both these factors had a significant impact in the differences between these methods, especially in the stroke lesion class. Training data was augmented by applying random flips to training patches, however we did not find this had a clear effect on results.

An important aspect to note is that WMH segmentation is notoriously challenging: for example, Bartko [25] and Anbeek et al. [13] consider similarity scores of 70 to be excellent, while Landis and Koch [145] consider scores of 40, 60 and 80 to be moderate, substantial and near perfect, respectively. With this in mind, we can consider average Dice scores for WMHs generated by the proposed uResNet, as well as those from DeepMedic and LPA to be substantial, while LGA generated only moderate results. It is important to note that LGA is an unsupervised method and that training data was only used to tune its κ parameter. Only uResNet and DeepMedic are capable of distinguishing between different types of lesions, and in this regard only uResNet produced an average stroke Dice score that could be considered moderate.

Part II

Dynamic MR Image Reconstruction

Chapter 6

Background: Dynamic MR Image Reconstruction

6.1 Introduction

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique which offers excellent spatial resolution and soft tissue contrast and is widely used for clinical diagnosis and research. Dynamic MRI attempts to reveal both spatial and temporal profiles of the underlying anatomy, which has a variety of applications such as cardiovascular imaging and perfusion imaging. However, the acquisition speed is fundamentally limited due to both hardware and physiological constraints as well as the requirement to satisfy the Nyquist sampling rate. Long acquisition time is not only a burden for patients but also makes MRI susceptible to motion artefacts.

In order to accelerate MRI acquisition, most approaches consider undersampling the data in k -space (frequency domain). Due to the violation of the Nyquist sampling theorem, undersampling introduces aliasing artefacts in the image domain. Images can be subsequently reconstructed by solving an optimisation problem that regularises the solution with assumptions on the underlying data, such as smoothness and sparsity. In dynamic MR imaging such as cardiac cine MRI, there exists significant correlations and redundancies between frames. Thus it is feasible to acquire only part of the k -space data, and then images can be reconstructed by

exploiting the spatial-temporal redundancies.

Problem Formulation

Let $\mathbf{x} \in \mathbb{C}^D$ denote a sequence of complex-valued MR images to be reconstructed, represented as a vector with $D = D_x D_y T$, and let $\mathbf{y} \in \mathbb{C}^M$ ($M \ll D$) represent the undersampled k -space measurements, where D_x and D_y are width and height of the frame respectively and T stands for the number of frames. Our problem is to reconstruct $\mathbf{x}$ from $\mathbf{y}$, which is commonly formulated as an unconstrained optimisation problem of the form:

$$\underset{\mathbf{x}}{\operatorname{argmin}} \quad \mathcal{R}(\mathbf{x}) + \lambda \|\mathbf{y} - \mathbf{F}_u \mathbf{x}\|_2^2 \quad (6.1)$$

Here $\mathbf{F}_u$ is an undersampling Fourier encoding matrix, $\mathcal{R}$ expresses regularisation terms on $\mathbf{x}$ and λ allows the adjustment of data fidelity based on the noise level of the acquired measurements $\mathbf{y}$. For compressed sensing (CS) and low-rank based approaches, the regularisation terms $\mathcal{R}$ often employed are ℓ_0 or ℓ_1 norms in the sparsifying domain of $\mathbf{x}$ as well as the rank or nuclear norm of $\mathbf{x}$ respectively.

6.2 Classical Methods for Dynamic MR Reconstruction

Over the years, a number of approaches have been proposed for the reconstruction of accelerated dynamic MR images. In general, these methods can be mainly divided into three categories, based on exploiting correlations in k -space, in time, and in both k -space and time [244].

Methods Exploiting k -space Correlations

The first class of approaches exploit the correlations between k -space points at the same time frame, and then reconstruct each frame independently from other time frames. This includes a variety of methods such as reduced field-of-view (FOV) [112] and parallel imaging methods

[98, 193]. These methods accelerate the MR acquisition by acquiring only part of the k -space samples, and then the missing data is recovered from the measured k -space data from the same time frame. This is based on the assumption that each k -space point contains information for other k -space points, and the missing information can be recovered by exploiting their correlations.

Methods Exploiting Temporal Correlations

The second group of strategies is to exploit redundancies in time, such as keyhole imaging [127] and data sharing [281]. In these methods, the missing data at a given position can be interpolated or extrapolated from the measured data at other time points, such as nearest-neighbor interpolation and linear interpolation. For instance, in view sharing approaches, one missing k -space point at a time frame is interpolated by the measured k -space points from neighboring frames and at the same position. This is based on the assumption that k -space samples at neighboring frames can capture similar information. In contrast to methods that exploit k -space correlations where each k -space point is reconstructed from other k -space points within the same frame, here each k -space point is reconstructed separately from other points across different time frames.

Methods Exploiting k - t Correlations

It is well known that in dynamic MRI there exists significant correlations in k -space and time. In order to increase the acquisition rate, most strategies have been designed to acquire part of the desired k - t measurements and then reconstruct the images by exploiting spatio-temporal redundancies within the data. Past literature has shown that exploiting spatio-temporal redundancy can greatly improve image reconstruction quality compared to compressed sensing (CS) based single frame reconstruction methods [129, 155].

One of the examples is the model-based k - t BLAST and k - t SENSE method [244], which take advantage of a-priori information about the x - f support obtained from the training stage and

then to remedy the aliasing artefacts during acquisition stage. An alternative popular approach is to exploit temporal redundancy to unravel from the aliasing by using CS approaches [44, 129] or CS combined with low-rank approaches [155, 184]. The class of methods which employ CS to the MRI reconstruction is termed as CS-MRI [163]. They assume that the image to be reconstructed has a sparse representation in a certain transform domain, and they need to balance sparsity in the transform domain against consistency with the acquired undersampled k -space data. For instance, an example of successful methods is k - t FOCUSS [129], which enforces the sparsity in x - f domain due to the periodic motions.

Specifically, consider a Cartesian k -space trajectory where k_x denotes the phase encoding direction, k_y denotes the readout direction, while $\sigma(x, t)$ denotes the image domain content at x and time t . The k -space measurement $v(k, t)$ is then formulated as:

$$v(k, t) = \int \sigma(x, t) e^{-j2\pi kx} dx = \int \int \rho(x, f) e^{-j2\pi(kx+ft)} dx df, \quad (6.2)$$

where $\rho(x, f)$ is the 2D spectral signal in x - f domain. This can also be represented in a matrix form:

$$\mathbf{v} = \mathcal{F}\rho, \quad (6.3)$$

in which $\mathbf{v}$ and ρ stand for the stacked k - t space measurement vectors and x - f image respectively, and $\mathcal{F}$ is the 2D Fourier transform along the x - f direction. From the perspective of compressed sensing, the problem can be formulated by exploiting the sparsity of the unknown signal:

$$\min \|\rho\|_1, \quad s.t. \|\mathbf{v} - \mathcal{F}\rho\|_2 \leq \epsilon, \quad (6.4)$$

where ϵ denotes the noise level. In k - t FOCUSS [128, 129], the underdetermined inverse problem was solved via a sparse reconstruction algorithm called FOCal Underdetermined System Solver (FOCUSS). The solution then can be expressed as the form that consists of a baseline signal $\bar{\rho}$ and its residual encoding for the n -th estimate of the x - f signal $\rho^{(n)}$:

$$\rho^{(n)} = \bar{\rho} + \Theta \mathcal{F}^H (\mathcal{F} \Theta \mathcal{F}^H + \lambda I)^{-1} (v - \mathcal{F} \bar{\rho}), \quad (6.5)$$

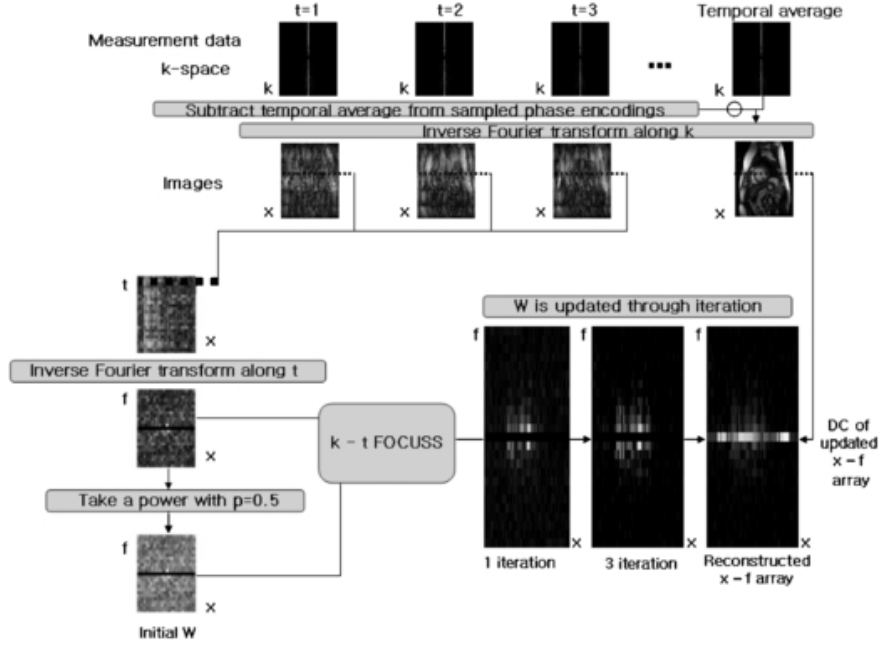


Figure 6.1: k-t FOCUSS reconstruction flow chart. Image adopted from [129].

where $\Theta = WW^H$, and W is updated by taking p power of the previous estimate ρ_{n-1} :

$$W_n = \begin{pmatrix} |\rho_{n-1}(1)|^p & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & |\rho_{n-1}(N)|^p \end{pmatrix}, \quad 1/2 \leq p \leq 1. \quad (6.6)$$

where $\rho_{n-1}(i)$ stands for the i -th element of the vector ρ_{n-1} . A flow chart of the k - t FOCUSS algorithm is shown in Fig. 6.1. The recovered signal is updated until convergence. For more details, please refer to [128, 129].

Besides, more recently, a low rank and sparse reconstruction scheme (k-t SLR) [155] introduces non-convex spectral norms and uses a spatio-temporal total variation norm in recovering the dynamic signal matrix. Dictionary learning approaches were also proposed to train an over-complete basis of atoms to optimally sparsify spatio-temporal data [44].

One of the main challenges associated with recovering an uncorrupted image is that both the undersampling strategy and a-priori knowledge of appropriate properties of the image need to be taken into account. These classical methods offer great potential for accelerated imaging,

however, there are a few challenges and limitations. Firstly, they often impose strong assumptions on the underlying data, requiring manual adjustments of hyperparameters depending on the application. For the optimisation based approaches, the regularisation functions and their hyper-parameters must be carefully selected, which are problem-specific and non-trivial. For example, over-imposing sparsity or ℓ_1 penalties can lead to cartoon-like/staircase artefacts. Secondly, it has been observed that these methods tend to result in blocky [101] and unnatural reconstructions. In addition, the reconstruction speeds of these methods are often slow due to requirement to solve iterative algorithms. Besides, these methods are not able to exploit the prior knowledge that can be learnt from the vast number of MRI exams routinely performed, which should be helpful to further guide the reconstruction process. Proposing a robust iterative algorithm is still an active area of research.

6.3 Deep Learning Methods for MR Reconstruction

Recently, deep learning-based MR reconstruction has gained popularity due to its promising results for solving inverse and compressed sensing problems. In comparison, unlike traditional approaches, the prior information and regularisation are learnt implicitly from data in deep learning methods, without having to specify them in the training objective. In general, current deep learning based approaches can be mainly divided into image-domain based methods which introduce improved regularisers and k -space based methods that focus on better interpolation using neural networks [140].

6.3.1 Deep Learning Methods in Image Space

For deep learning approaches developed in image space, two paradigms have emerged: the first class of approaches proposes to use convolutional neural networks (CNNs) to learn an end-to-end mapping, where architectures such as SRCNN [73] or U-net [206] are often chosen for MR image reconstruction [103, 152, 259, 260]. The second class of approaches attempts to make

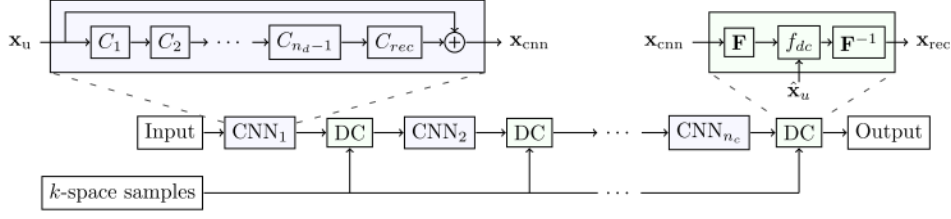


Figure 6.2: Illustration of DC-CNN architecture. Image taken from [214].

each stage of iterative optimisation learnable by unrolling the end-to-end pipeline into a deep network [4, 5, 101, 214, 235].

The second group of methods extend naturally from Eq. 6.1, where each layer in the network corresponds to an iteration step of a classical algorithm. For instance, Hammernik et al. [101] introduced a trainable formulation for accelerated parallel imaging (PI) based MRI reconstruction termed variational network, which follows the structure of variational methods and gradient-based optimisation. ADMM-Net [235] was proposed by reformulating an alternating direction method of multipliers (ADMM) algorithm to a deep network, where each stage of the architecture corresponds to an iteration in the ADMM algorithm. More recently, Schlemper et al. [214] proposed a cascade network (DC-CNN) which simulated the iterative reconstruction of dictionary learning-based methods. An illustration of the DC-CNN approach is shown in Fig. 6.2, where it formulates the closed-form solution in k -space as a data consistency layer with the form:

$$f_{dc}(\mathbf{x}; \mathbf{y}, \lambda_0, \Omega) = \mathbf{F}^H \mathbf{\Lambda} \mathbf{F} \mathbf{x} + \frac{\lambda_0}{1+\lambda_0} \mathbf{F}_u^H \mathbf{y},$$

$$\mathbf{\Lambda}_{kk} = \begin{cases} 1 & \text{if } k \notin \Omega \\ \frac{1}{1+\lambda_0} & \text{if } k \in \Omega \end{cases} \quad (6.7)$$

in which $\mathbf{F}$ is the full Fourier encoding matrix (a discrete Fourier transform in this case), Ω is an index set of the acquired k -space samples and $\mathbf{\Lambda}$ is a diagonal matrix. In the noiseless setting ($\lambda_0 \rightarrow \infty$), the k -th predicted coefficient is simply replaced by the original coefficient if it has been sampled. Similarly, Aggarwal et al. [7] proposed to share the same set of parameters across each stage of reconstructions, with the aim of reducing the needed parameters, and a conjugate gradient step was employed as data consistency.

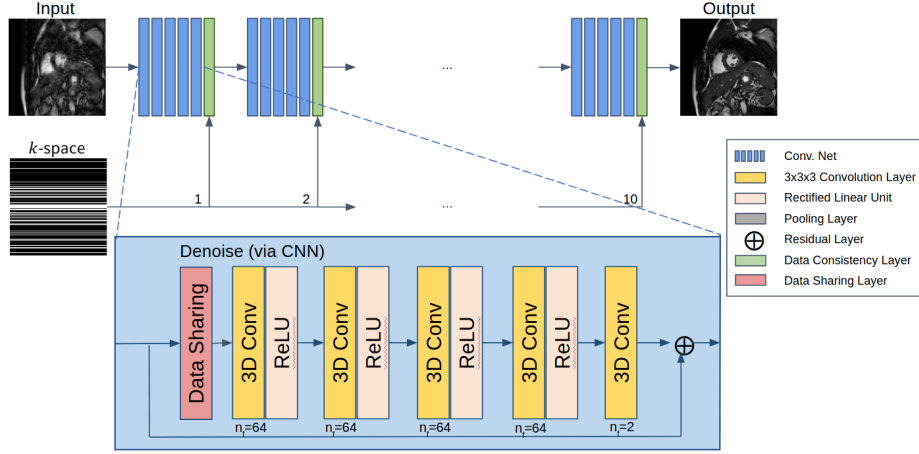


Figure 6.3: Illustration of DC-CNN architecture for dynamic MR reconstruction. Image adopted from [215].

Most approaches so far have focused on 2D images, whereas only a few approaches exist for dynamic MR reconstruction [26, 215]. An example of dynamic MR reconstruction is an extension of DC-CNN [215], as shown in Fig. 6.3. Here 3D convolutions were employed to exploit information contained in both spatial and temporal dimension. In addition, a data sharing (DS) layer was proposed to approximate missing k -space samples from adjacent frames, with the assumption that neighboring k -space samples often capture similar information. A schematic illustration of the data sharing scheme is shown in Fig. 6.4, where for each frame t , frames from $t - n_{adj}$ to $t + n_{adj}$ are considered to fill missing k -space samples for frame t by averaging. While they show promising results, the optimal architecture, training scheme and configuration spaces are yet to be fully explored.

6.3.2 Deep Learning Methods in k -space

Recently, there has been a lot of interests in reconstructing MR images in k -space, via using neural networks to improve the k -space interpolation techniques. These approaches can be mainly divided into two groups: the first class of methods rely on scan-specific autocalibrating signals (ACS) to estimate the interpolation kernels by training a neural network, while the second class of approaches utilises training databases to fill the missing k -space samples.

A representative approach for the scan-specific deep learning methods is robust artificial-

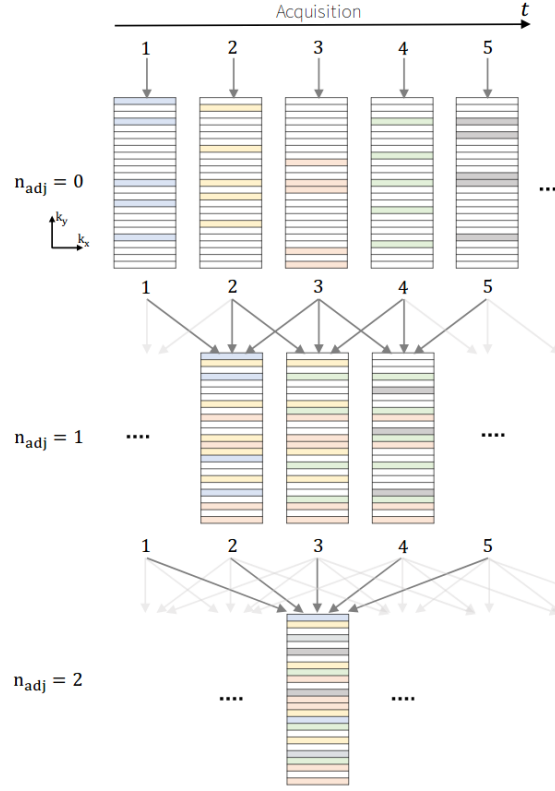


Figure 6.4: Schematic illustration of data sharing layer. The acquired k -space lines are color encoded. Missing entries in each frame are aggregated using values from n_{adj} neighboring frames. Overlapped lines are averaged. Image adopted from [215].

neural-networks for k -space interpolation (RAKI) [10]. It used a CNN network to interpolate missing k -space samples from acquired ones, and network parameters were learnt by training the network on the ACS data with an MSE loss function. One advantage of such scan-specific methods is that they do not require any training databases, and can be applicable to cases where fully-sampled ground truth data are difficult to acquire. However, network needs to be trained for each scan independently and it also requires more calibration data. Based on this approach, a residual RAKI (rRAKI) has also been proposed, where a CNN learns to remove artefacts based on the residual derived from GRAPPA interpolation kernel.

In contrast, the second group of approaches learn to fill the missing k -space data from large training databases. For instance, Han et al. [102] proposed to employ a CNN network for a Hankel matrix based approaches in k -space to complete the k -space data. Specifically, a CNN was used to replace the k -space completion in ALOHA method, where it showed that CNN

based method can achieve better performance compared to the original ALOHA approach. Some other works such as DeepSPIRiT and Zhang et al. [280] also explored the k -space interpolation using neural networks for parallel imaging by training on a large group of training data.

In addition, some approaches proposed to reconstruct MR image by exploiting the cross-domain knowledge: in both k -space and image domain. One example of such methods is the KIKI-net [78], where it learnt the reconstruction network in both k -space and image space iteratively, with data consistency layer interleaved at each stage. It showed that by exploiting information from complementary domains simultaneously, such approaches can result in better reconstruction quality than methods training on each domain separately.

However, so far only a handful of approaches exist [26, 215] for dynamic reconstruction. Hence, the applicability of deep learning models to this problem is yet to be fully explored. In addition, many proposed deep learning architectures are often generic and are not optimised for specific applications. In particular, a core question for dynamic reconstruction is how to optimally exploit spatio-temporal redundancy. It will be interesting to design a network architecture and regulate the mechanics of network layers to efficiently learn such spatio-temporal data representation.

6.4 Conclusions

This chapter gives a brief overview of the background and related work in dynamic MR image reconstruction. In particular, this chapter reviewed some of the representative classical methods based on their exploiting correlations in k -space, in time and in both k space and time. More recent deep learning models for MR image reconstruction have also been presented, according to their reconstruction in image domain or in k -space. In the remaining chapters of this part, two novel deep learning models will be introduced, which exploit spatio-temporal redundancies for dynamic cardiac MR image reconstruction.

Chapter 7

Convolutional Recurrent Neural Network for Dynamic MRI Reconstruction

This chapter is based on:

- **C. Qin**, J. Schlemper, J. Caballero, J. Hajnal, A. Price and D. Rueckert. Convolutional Recurrent Neural Networks for Dynamic MR Image Reconstruction. *IEEE Transactions on Medical Imaging*, 38(1), 280-290, 2019.

This work is contributed equally with J. Schlemper.

7.1 Introduction

Accelerating the data acquisition of dynamic MRI leads to a challenging ill-posed inverse problem, which has received great interest from both the signal processing and machine learning communities over the last decades. The key ingredient to the problem is how to exploit the temporal correlations of the MR sequence to resolve aliasing artefacts. Traditionally, such

observation led to a formulation of an optimisation problem, which was solved using iterative algorithms. Recently, however, deep learning based-approaches have gained significant popularity due to their ability to solve general inverse problems. In this work, we propose a unique, novel convolutional recurrent neural network (CRNN) architecture which reconstructs high quality cardiac MR images from highly undersampled k -space data by jointly exploiting the dependencies of the temporal sequences as well as the iterative nature of the traditional optimisation algorithms. The proposed method is termed as *CRNN-MRI*.

Specifically, in the work, we firstly formulate a general optimisation problem for solving accelerated dynamic MRI based on variable splitting and alternate minimisation. We then show how this algorithm can be seen as a network architecture. In particular, the proposed method consists of a CRNN block which acts as the proximal operator and a data consistency layer corresponding to the classical data fidelity term. In addition, the CRNN block employs recurrent connections across each iteration step, allowing reconstruction information to be shared across the multiple iterations of the process. Secondly, we incorporate bidirectional convolutional recurrent units evolving over time to exploit the temporal dependency of the dynamic sequences and effectively propagate the contextual information across time frames of the input. As a consequence, the unique CRNN architecture jointly learns representations in a recurrent fashion evolving over both *time sequences* as well as *iterations* of the reconstruction process, effectively combining the benefits of traditional iterative methods and deep learning.

The contributions of this work are the following: Firstly, the optimisation problem of dynamic data is viewed as a recurrent network and a novel CRNN architecture which simultaneously incorporates the recurrence existing in both temporal and iteration sequential steps is described. Secondly, we demonstrate that the proposed method shows promising results and improves upon the current state-of-the-art dynamic MR reconstruction methods both in reconstruction accuracy and speed. Finally, the proposed architecture is compared to 3D CNN which does not impose the recurrent structure. The proposed method is shown to outperform the CNN at different undersampling rates and speed, while requiring significantly fewer parameters.

7.2 Method

7.2.1 Problem Formulation

As shown in Section 6.1, the MR reconstruction problem can be formulated as an unconstrained optimisation problem of the form as described in Eq. 6.1. Here we present the formulation again:

$$\operatorname{argmin}_{\mathbf{x}} \mathcal{R}(\mathbf{x}) + \lambda \|\mathbf{y} - \mathbf{F}_u \mathbf{x}\|_2^2 \quad (7.1)$$

In general, Eq. 7.1 is a non-convex function and hence, the variable splitting technique is usually adopted to decouple the fidelity term and the regularisation term. By introducing an auxiliary variable $\mathbf{z}$ that is constrained to be equal to $\mathbf{x}$, Eq. 7.1 can be reformulated to minimise the following cost function via the penalty method:

$$\operatorname{argmin}_{\mathbf{x}, \mathbf{z}} \mathcal{R}(\mathbf{z}) + \lambda \|\mathbf{y} - \mathbf{F}_u \mathbf{x}\|_2^2 + \mu \|\mathbf{x} - \mathbf{z}\|_2^2 \quad (7.2)$$

where μ is a penalty parameter. By applying alternate minimisation over $\mathbf{x}$ and $\mathbf{z}$, Eq. 7.2 can be solved via the following iterative procedures:

$$\mathbf{z}^{(i+1)} = \operatorname{argmin}_{\mathbf{z}} \mathcal{R}(\mathbf{z}) + \mu \|\mathbf{x}^{(i)} - \mathbf{z}\|_2^2 \quad (7.3a)$$

$$\mathbf{x}^{(i+1)} = \operatorname{argmin}_{\mathbf{x}} \lambda \|\mathbf{y} - \mathbf{F}_u \mathbf{x}\|_2^2 + \mu \|\mathbf{x} - \mathbf{z}^{(i+1)}\|_2^2 \quad (7.3b)$$

where $\mathbf{x}^{(0)} = \mathbf{x}_u = \mathbf{F}_u^H \mathbf{y}$ is the zero-filled reconstruction taken as an initialisation and $\mathbf{z}$ can be seen as an intermediate state of the optimisation process. For MRI reconstruction, Eq. 7.3b is often regarded as a *data consistency* (DC) step where a closed-form solution [214] can be given as in Eq. 6.7. Please refer to [214] for more details of formulating Eq. 6.7 as a data consistency layer in a neural network. Eq. 7.3a is the proximal operator of the prior $\mathcal{R}$, and instead of explicitly determining the form of the regularisation term, we propose to directly learn the proximal operator by using a convolutional recurrent neural network (CRNN).

Previous deep learning approaches such as Deep-ADMM net [235] and method proposed by

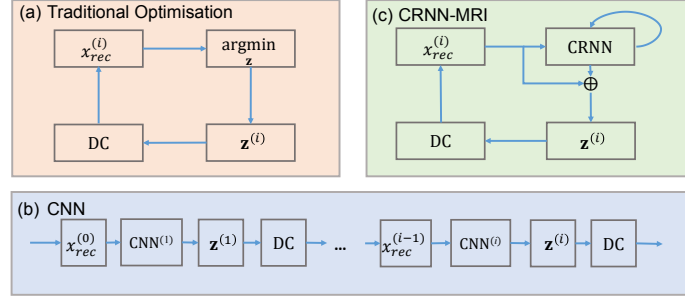


Figure 7.1: (a) Traditional optimisation algorithm using variable splitting and alternate minimisation approach, (b) the optimisation unrolled into a deep convolutional network incorporating the data consistency step, and (c) the proposed architecture which models optimisation recurrence.

Schlemper et al. [214] unroll the traditional optimisation algorithm. Hence, their models learn a sequence of transition $\mathbf{x}^{(0)} \rightarrow \mathbf{z}^{(1)} \rightarrow \mathbf{x}^{(1)} \rightarrow \dots \rightarrow \mathbf{z}^{(N)} \rightarrow \mathbf{x}^{(N)}$ to reconstruct the image, where each state transition at stage (i) is an operation such as convolutions independently parameterised by $\boldsymbol{\theta}$, nonlinearities or a data consistency step. However, since the network implicitly learns some form of proximal operator at each iteration, it may be redundant to individually parameterise each step. In our formulation, we model each optimisation stage (i) as a learnt, *recurrent*, forward encoding step $f_i(\mathbf{x}^{(i-1)}; \boldsymbol{\theta}, \mathbf{y}, \lambda, \Omega)$. The difference is that now we use one model which performs proximal operator, however, it also allows itself to propagate information across iteration, making it adaptable for the changes across the optimisation steps. The detail will be discussed in the following section. The different strategies are illustrated in Fig. 7.1.

7.2.2 CRNN for MRI reconstruction

RNN is a class of neural networks that makes use of sequential information to process sequences of inputs. They maintain an internal state of the network acting as a “memory”, which allows RNNs to naturally lend themselves to the processing of sequential data. Inspired by iterative optimisation schemes of Eq. 7.3, we propose a novel convolutional RNN (CRNN) network. In

where CRNN is a learnable block explained hereafter, DC is the data consistency step treated as a network layer, $\mathbf{x}_{rec}^{(i)}$ is the progressive reconstruction of the undersampled image $\mathbf{x}_u$ at iteration i with $\mathbf{x}_{rec}^{(0)} = \mathbf{x}_u$, $\mathbf{x}_{rnn}^{(i)}$ is the intermediate reconstruction image before the DC layer, and $\mathbf{y}$ is the acquired k -space samples. Note that the variables $\mathbf{x}_{rec}$, $\mathbf{x}_{rnn}$ are analogous to $\mathbf{x}$, $\mathbf{z}$ in Eq. 7.3 respectively. Here, we use CRNN to encode the update step, which can be seen as one step of a gradient descent in the sense of objective minimisation, or a more general approximation function regressing the difference $\mathbf{z}^{(i+1)} - \mathbf{x}^{(i)}$, i.e. the distance required to move to the next state. Moreover, note that in every iteration, CRNN updates its internal state $\mathcal{H}$ given an input which is discussed shortly. As such, CRNN also allows information to be propagated efficiently across iterations, in contrast to the sequential models using CNNs which collapse the intermediate feature representation to $\mathbf{z}^{(i)}$.

In order to exploit the dynamic nature and the temporal redundancy of our data, we further propose to jointly model the recurrence evolving over time for dynamic MRI reconstruction. The proposed CRNN-MRI network and CRNN block are shown in Fig. 7.2(a), in which CRNN block is comprised of 5 components:

- bidirectional convolutional recurrent units evolving over time and iterations (BCRNN-t-i),
- convolutional recurrent units evolving over iterations only (CRNN-i),
- 2D convolutional neural network (CNN),
- residual connection and
- DC layers.

Details of the components of the proposed network will be introduced in the following.

CRNN-i

As aforementioned, the iterative optimisation procedures is explicitly encapsulated with RNNs. In the CRNN-i unit, the iteration step is viewed as the sequential step in the vanilla RNN. If

the network is unfolded over the iteration dimension, the network can be illustrated as in Fig. 7.2(b), where information is propagated between iterations. Here we use $\mathbf{H}$ to denote the feature representation of our sequence of frames throughout the network. $\mathbf{H}_l^{(i)}$ denotes the representation at layer l (subscript) and iteration step i (superscript). Therefore, at iteration (i) , given the input $\mathbf{H}_{l-1}^{(i)}$ and the previous iteration's hidden state $\mathbf{H}_l^{(i-1)}$, the hidden state $\mathbf{H}_l^{(i)}$ at layer l of a CRNN-i unit can be formulated as:

$$\mathbf{H}_l^{(i)} = \sigma(\mathbf{W}_l * \mathbf{H}_{l-1}^{(i)} + \mathbf{W}_i * \mathbf{H}_l^{(i-1)} + \mathbf{B}_l). \quad (7.6)$$

Here $*$ represents convolution operation, $\mathbf{W}_l$ and $\mathbf{W}_i$ represent the filters of input-to-hidden convolutions and hidden-to-hidden recurrent convolutions evolving over iterations respectively, and $\mathbf{B}_l$ represents a bias term. Here $\mathbf{H}_l^{(i)}$ is the representation of the whole T sequence with shape $(batchsize, T, n_c, D_x, D_y)$, where n_c is the number of channels which is 2 at the input and output but is greater while processing inside the network, and the convolutions are computed on the last two dimensions. The latent features are activated by the rectified linear unit (ReLU) as a choice of nonlinearity, i.e. $\sigma(x) = \max(0, x)$.

The CRNN-i unit offers several advantages compared to independently unrolling convolutional filters at each stage. Firstly, compared to CNNs where the latent representation from the previous state is not propagated, the hidden-to-hidden iteration connections in CRNN-i units allow contextual spatial information gathered at previous iterations to be passed to the future iterations. This enables the reconstruction step at each iteration to be optimised not only based on the output image but also based on the hidden features from previous iterations, where the hidden connection convolutions can “memorise” the useful features to avoid redundant computation. Secondly, as the iteration number increases, the effective receptive field of a CRNN-i unit in the spatial domain also expands whereas CNN resets it at each iteration. This property allows the network to further improve the reconstruction by allowing it to have better contextual support. In addition, since the weight parameters are shared across iterations, it greatly reduces the number of parameters compared to CNNs, potentially offering better generalisation properties.

BCRNN-t-i

Dynamic MR images exhibit high temporal redundancy, which is often exploited as a-priori knowledge to regularise the reconstruction. Hence, it is also beneficial for the network to learn the dynamics of sequences. To this extent, we propose a bidirectional convolutional recurrent unit (BCRNN-t-i) to exploit both temporal *and* iteration dependencies jointly. BCRNN-t-i includes three convolution layers: one on the input which comes into the unit from the previous layer indicated by the green arrows in Fig. 7.2(c), one on the hidden state from the past and future time frames as shown by the red arrows, and the one on the hidden state from the previous iteration of the unit (blue arrows in Fig. 7.2(c)). Note that we simultaneously consider temporal dependencies from past and future time frames, and the encoding weights are shared for both directions due to the cyclic cardiac motion. The output for the BCRNN-t-i layer is obtained by summing the feature maps learnt from both directions. The illustration figure of the unit when it is unfolded over time sequence is shown in Fig. 7.2(c).

As information along temporal dimensions needs to be propagated in this unit, here an additional index t in the notation is introduced to represent the variables related with time frame t . Here $\mathbf{H}_{l,t}^{(i)}$ represents feature representations at l -th layer, time frame t , and at iteration i , $\vec{\mathbf{H}}_{l,t}^{(i)}$ denotes the representations calculated when information is propagated forward inside the BCRNN-t-i unit, and similarly, $\overleftarrow{\mathbf{H}}_{l,t}^{(i)}$ denotes the one in the backward direction. Therefore, for the formulation of BCRNN-t-i unit, given (1) the current input representation of the l -th layer at time frame t and iteration step i , which is the output representation from $(l-1)$ -th layer $\mathbf{H}_{l-1,t}^{(i)}$, (2) the previous iteration's hidden representation within the same layer $\mathbf{H}_{l,t}^{(i-1)}$, (3) the hidden representation of the past time frame $\vec{\mathbf{H}}_{l,t-1}^{(i)}$, and the hidden representation of the future time frame $\overleftarrow{\mathbf{H}}_{l,t+1}^{(i)}$, then the hidden state representation of the current l -th layer of time frame t at iteration i , $\mathbf{H}_{l,t}^{(i)}$ with shape $(batchsize, n_c, D_x, D_y)$, can be formulated as:

$$\begin{aligned}\vec{\mathbf{H}}_{l,t}^{(i)} &= \sigma(\mathbf{W}_l * \mathbf{H}_{l-1,t}^{(i)} + \mathbf{W}_t * \vec{\mathbf{H}}_{l,t-1}^{(i)} + \mathbf{W}_i * \mathbf{H}_{l,t}^{(i-1)} + \vec{\mathbf{B}}_l), \\ \overleftarrow{\mathbf{H}}_{l,t}^{(i)} &= \sigma(\mathbf{W}_l * \mathbf{H}_{l-1,t}^{(i)} + \mathbf{W}_t * \overleftarrow{\mathbf{H}}_{l,t+1}^{(i)} + \mathbf{W}_i * \mathbf{H}_{l,t}^{(i-1)} + \overleftarrow{\mathbf{B}}_l), \\ \mathbf{H}_{l,t}^{(i)} &= \vec{\mathbf{H}}_{l,t}^{(i)} + \overleftarrow{\mathbf{H}}_{l,t}^{(i)},\end{aligned}\tag{7.7}$$

Similar to the notation in Section 7.2.2, $\mathbf{W}_t$ represents the filters of recurrent convolutions evolving over time. When $l = 1$ and $i = 1$, $\mathbf{H}_{0,t}^{(1)} = \mathbf{x}_{u_t}$, that is the t -th frame of undersampled input data, and when $l = 1$ and $i = 2, \dots, T$, $\mathbf{H}_{0,t}^{(i)} = \mathbf{x}_{rec_t}^{(i-1)}$, which stands for the t -th frame of the intermediate reconstruction result from iteration $i - 1$. For $\mathbf{H}_{l,t}^{(0)}$, $\vec{\mathbf{H}}_{l,0}^{(i)}$ and $\overleftarrow{\mathbf{H}}_{l,T+1}^{(i)}$, they are set to be zero initial hidden states.

The temporal connections of BCRNN-t-i allow information to be propagated across the whole T time frames, enabling it to learn the differences and correlations of successive frames. The filter responses of recurrent convolutions evolving over time express dynamic changing biases, which focus on modelling the temporal changes across frames, while the filter responses of recurrent convolutions over iterations focus on learning the spatial refinement across consecutive iteration steps. In addition, we note that learning recurrent layers along the temporal direction is different from using 3D convolution along the space and temporal direction. 3D convolution seeks invariant features across space-time, hence several layers of 3D convolutions are required before the information from the whole sequence can be propagated to a particular time frame. On the other hand, learning recurrent 2D convolutions enables the model to easily and efficiently propagate the information through time, which also yields fewer parameters and a lower computational cost.

In summary, the set of hidden states for a CRNN block to update at iteration i is $\mathcal{H} = \{\mathbf{H}_l^{(i)}, \mathbf{H}_{l,t}^{(i)}, \overleftarrow{\mathbf{H}}_{l,t}^{(i)}, \vec{\mathbf{H}}_{l,t}^{(i)}\}$, for $l = 1, \dots, L$ and $t = 1, \dots, T$, where L is the total number of layers in the CRNN block and T is the total number of time frames.

7.2.3 Network Learning

Given the training data S of input-target pairs $(\mathbf{x}_u, \mathbf{x}_t)$, the network learning proceeds by minimising the pixel-wise mean squared error (MSE) between the predicted reconstructed MR image and the fully sampled ground truth data:

$$\mathcal{L}(\theta) = \frac{1}{n_S} \sum_{(\mathbf{x}_u, \mathbf{x}_t) \in S} \|\mathbf{x}_t - \mathbf{x}_{rec}\|_2^2 \quad (7.8)$$

where $\boldsymbol{\theta} = \{\mathbf{W}_l, \mathbf{W}_i, \mathbf{W}_t, \mathbf{B}_l\}$, $l = 1 \dots L$, and n_S stands for the number of samples in the training set S . Note that the total number of time sequences T and iteration steps N assumed by the network before performing the reconstruction are free parameters that must be specified in advance. The network weights were initialised using He initialization [106] and they were trained using the Adam optimiser [138]. During training, gradients were hard-clipped to the range of $[-5, 5]$ to mitigate the gradient explosion problem. The network was implemented in Python using Theano and Lasagne libraries.

7.3 Experiments

7.3.1 Dataset and Implementation Details

The proposed method was evaluated using a complex-valued MR dataset consisting of 10 fully sampled short-axis cardiac cine MR scans. Each scan contains a single slice SSFP acquisition with 30 temporal frames, which have a 320×320 mm field of view and 10 mm thickness. The raw data consists of 32-channel data with sampling matrix size 192×190 , which was then zero-filled to the matrix size 256×256 . The raw multi-coil data was reconstructed using SENSE [193] with no undersampling and retrospective gating. Coil sensitivity maps were normalised to a body coil image and used to produce a single complex-valued reconstructed image. In experiments, the complex valued images were back-transformed to regenerate k -space samples, simulating a fully sampled single-coil acquisition. The input undersampled image sequences were generated by randomly undersampling the k -space samples using Cartesian undersampling masks, with undersampling patterns adopted from [129]: for each frame the eight lowest spatial frequencies were acquired, and the sampling probability of k -space lines along the phase-encoding direction was determined by a zero-mean Gaussian distribution. Note that the undersampling rates are stated with respect to the matrix size of raw data, which is 192×190 .

The architecture of the proposed network used in the experiment is shown in Fig. 7.2 which inherits similar overall structure as in [215] but with incorporation of the proposed new units.

Specifically, each iteration of the CRNN block contains five units: one layer of BCRNN-t-i, followed by three layers of CRNN-i units, and followed by a CNN unit. The structure of CRNN-i and BCRNN-t-i units follow the descriptions in Section 7.2.2 based on vanilla RNNs [77], with hidden-to-hidden connections between iteration steps as well as temporal sequences. We used vanilla RNNs [77] here to model the recurrence due to its simplicity and thus for the proof of concept. Note this can be naturally generalised to other RNN units, such as long short-term memory (LSTM) and gated recurrent unit (GRU), which are considered to have better memory properties, although using these units would significantly increase computational complexity. In addition, this architecture is by no means optimal and more layers can be added to increase the ability of our network to better capture the data structures (see Section 7.3.4 for comparisons).

In details, for all CRNN-i and BCRNN-t-i units, we used a kernel size $k = 3$ and the number of filters was set to $n_f = 64$ for Proposed-A and $n_f = 128$ for Proposed-B in Table 7.1. The CNN after the CRNN-i units contains one convolution layer with $k = 3$ and $n_f = 2$, which projects the extracted representation back to the image domain which contains complex-valued images expressed using two channels. For all convolutional layers, we used stride = 1 and paddings with half the filter size (rounded down) on both size. The output of the CRNN block is connected to the residual connection, which sums the output of the block with its input. Finally, we used DC layers on top of the CRNN output layers. During training, the iteration step is set to be $N = 10$, and the time sequence for training is $T = 30$.

The evaluation was done via a 3-fold cross validation, where for two folds we train on 7 subjects then test on 3 subjects, and for the remaining fold we train on 6 subjects and test on 4 subjects. While the original sequence has size $256 \times 256 \times T$, for the training, we extract patches of size $256 \times D_{patch} \times T$, where $D_{patch} = 32$ is the patch size and the direction of patch extraction corresponds to the frequency-encoding direction. Note that since we only consider Cartesian undersampling, the aliasing occurs only along the phase encoding direction, so patch extraction does not alter the aliasing artefact. Patch extraction as well as data augmentation was performed on-the-fly, with random affine and elastic transformations on the image data. For pre-processing, all the image intensities were normalised to the range of $[0, 1]$, and evaluations were performed on the normalised images. Undersampling masks were also generated randomly

following patterns in [129] for each input. During test time, the network trained on patches is directly applied on the whole sequence of the original image. The minibatch size during the training was set to 1, and we observed that the performance can reach a plateau within 6×10^4 backpropagations.

7.3.2 Evaluation Method

We compared the proposed method with the representative algorithms of the CS-based dynamic MRI reconstruction, such as k-t FOCUSS [129] and k-t SLR [155], and two variants of 3D CNN networks named 3D CNN-S and 3D CNN in our experiments. The built baseline 3D CNN networks share the same architecture with the proposed CRNN-MRI network but all the recurrent units and 2D CNN units were replaced with 3D convolutional units, that is, in each iteration, the 3D CNN block contain 5 layers of 3D convolutions, one DC layer and a residual connection. Here 3D CNN-S refers to network sharing weights across iterations, however, this does not employ the hidden-to-hidden connection as in the CRNN-i unit. The 3D CNN-S architecture was chosen so as to make a fair comparison with the proposed model using a comparable number of network parameters. In contrast, 3D CNN refers to the network without weight sharing, in which the network capacity is $N = 10$ times of that of 3D CNN-S, and approximately 12 times more than that of our proposed method (Proposed-A). For the 3D CNN approaches, the receptive field size is $11 \times 11 \times 11$, as the receptive field size is “reset” after each data consistency layer. In contrast, for the proposed method, due to the hidden connections between iterations and bidirectional temporal connections, by tracing the longest path of the convolution layers involved in the forward pass, including both temporal and iterative directions, in theory, the receptive field size is $309 \times 309 \times 30$ (154 layers of CNNs for the middle frame in a sequence of 30 frames). However, the network still may predominantly relies on local features coming from the partial reconstruction. Nevertheless, the RNN has the ability to exploit the features with larger filter size if needed, which is not the case for 3D CNNs.

Reconstruction results were evaluated based on the following quantitative metrics: MSE,

Table 7.1: Performance comparisons (MSE, PSNR:dB, SSIM, and HFEN) on dynamic cardiac data with different acceleration rates. MSE is scaled to 10^{-3} . The bold numbers are better results of the proposed methods than that of the other methods.

Method	k-t FOCUSS	k-t SLR	3D CNN-S	3D CNN	Proposed-A	Proposed-B	
Capacity	-	-	338,946	3,389,460	262,020	1,040,132	
6×	MSE	0.592 (0.199)	0.371(0.155)	0.385 (0.124)	0.275 (0.096)	0.261 (0.097)	0.201 (0.074)
	PSNR	32.506 (1.516)	34.632 (1.761)	34.370 (1.526)	35.841 (1.470)	36.096 (1.539)	37.230 (1.559)
	SSIM	0.953 (0.040)	0.970 (0.033)	0.976 (0.008)	0.983 (0.005)	0.985 (0.004)	0.988 (0.003)
	HFEN	0.211 (0.021)	0.161 (0.016)	0.170 (0.009)	0.138 (0.013)	0.131 (0.013)	0.112 (0.010)
9×	MSE	1.234 (0.801)	0.846 (0.572)	0.929 (0.474)	0.605 (0.324)	0.516 (0.255)	0.405 (0.206)
	PSNR	29.721 (2.339)	31.409 (2.404)	30.838 (2.246)	32.694 (2.179)	33.281 (1.912)	34.379 (2.017)
	SSIM	0.922 (0.043)	0.951 (0.025)	0.950 (0.016)	0.968 (0.010)	0.972 (0.009)	0.979 (0.007)
	HFEN	0.310(0.041)	0.260 (0.034)	0.280 (0.034)	0.215 (0.021)	0.201 (0.025)	0.173 (0.021)
11×	MSE	1.909 (0.828)	1.237 (0.620)	1.472 (0.733)	0.742 (0.325)	0.688 (0.290)	0.610 (0.300)
	PSNR	27.593 (2.038)	29.577 (2.211)	28.803 (2.151)	31.695 (1.985)	31.986 (1.885)	32.575 (1.987)
	SSIM	0.880 (0.060)	0.924 (0.034)	0.925 (0.022)	0.960 (0.010)	0.964 (0.009)	0.968 (0.011)
	HFEN	0.390 (0.023)	0.327 (0.028)	0.363 (0.041)	0.257 (0.029)	0.248 (0.033)	0.227 (0.030)
Time	15s	451s	8s	8s	3s	6s	

peak-to-noise-ratio (PSNR), structural similarity index (SSIM) [263] and high frequency error norm (HFEN) [200]. The choice of the these metrics was made to evaluate the reconstruction results with complimentary emphasis. MSE and PSNR were chosen to evaluate the overall accuracy of the reconstruction quality. SSIM put emphasis on image quality perception. HFEN was used to quantify the quality of the fine features and edges in the reconstructions, and here we employed the same filter specification as in [178, 200] with the filter kernel size 15×15 pixels and a standard deviation of 1.5 pixels. For PSNR and SSIM, it is the higher the better, while for MSE and HFEN, it is the lower the better.

7.3.3 Results

The comparison results of all methods are reported in Table 7.1, where we evaluated the quantitative metrics, network capacity and reconstruction time. Numbers shown in Table 7.1 are mean values of corresponding metrics with standard deviation of different subjects in parenthesis. Bold numbers in Table 7.1 indicate the better performance of the proposed methods than the competing ones. Compared with the baseline method (k-t FOCUSS and k-t SLR), the proposed methods outperform them by a considerable margin at different acceleration rates. When compared with deep learning methods, note that the network capacity of Proposed-

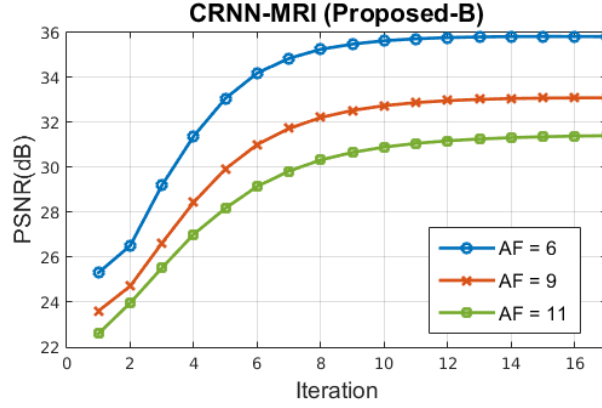


Figure 7.3: Mean PSNR values (Proposed-B) vary with the number of iterations at test time on data with different acceleration factors. Here AF stands for acceleration factor.

A is comparable with that of 3D CNN-S and the capacity of Proposed-B is around one third of that of 3D CNN. Though their capacities are much smaller, both Proposed-A and Proposed-B outperform 3D CNN-S and 3D CNN for all acceleration rates by a large margin, which shows the competitiveness and effectiveness of our method. In addition, we can see a substantial improvement of the reconstruction results on all acceleration rates and in all metrics when the number of network parameters is increased for the proposed method (Proposed-B), and therefore we will only show the results from Proposed-B in the following. The number of iterations used by the network at test time is set to be the same as the training stage, which is $N = 10$, however, if the iteration number is increased up to $N = 17$, it shows an improvement of 0.324dB on average. Fig. 7.3 shows the model's performance varying with the number of iterations at test time. Similarly, visualisation results of intermediate steps during the iterations of a reconstruction from $9\times$ undersampling data are shown in Fig. 7.4, where we can observe the gradual improvement of the reconstruction quality from iteration step 1 to 10, which is consistent with the quantitative results as in Fig. 7.3.

A comparison of the visualisation results of a reconstruction from $9\times$ acceleration is shown in Fig. 7.5 with the reconstructed images and their corresponding error maps from different reconstruction methods. As one can see, our proposed model (Proposed-B) can produce more faithful reconstructions for those parts of the image around the myocardium where there are large temporal changes. This is reflected by the fact that RNNs effectively use a larger receptive

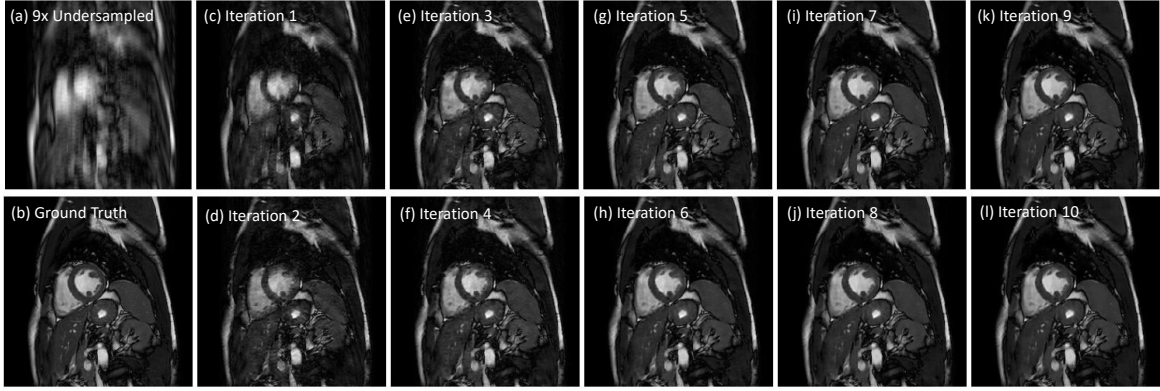


Figure 7.4: Visualisation results of intermediate steps during the iterations of a reconstruction. (a) Undersampled image by acceleration factor 9 (b) Ground Truth (c-l) Results from intermediate steps 1 to 10 in a reconstruction process.

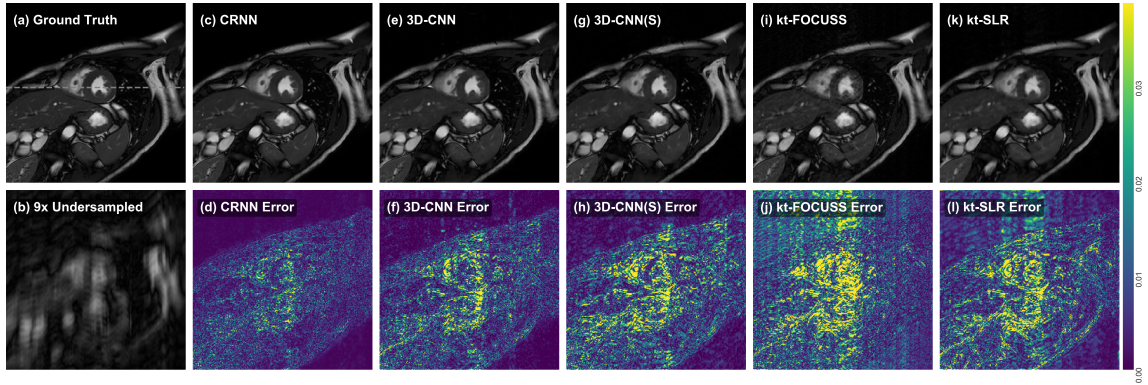


Figure 7.5: The comparison of reconstructions on spatial dimension with their error maps. (a) Ground Truth (b) Undersampled image by acceleration factor 9 (c,d) Proposed-B (e,f) 3D CNN (g,h) 3D CNN-S (i,j) k-t FOCUSS (k,l) k-t SLR

field to capture the characteristics of aliasing seen within the anatomy. Their temporal profiles at $x = 120$ are shown in Fig. 7.6. Similarly, one can see that the proposed model has overall much smaller error, faithfully modelling the dynamic data. It could be due to the fact that spatial and temporal features are learnt separately in the proposed model while 3D CNN seeks invariant feature learning across space and time.

In terms of speed, the proposed RNN-based reconstruction is faster than the 3D CNN approaches because it only performs convolution along time once per iteration, removing the redundant 3D convolutions which are computationally expensive. Reconstruction time of 3D CNN and the proposed methods reported in Table 7.1 were calculated on a GPU GeForce GTX 1080, and the time for k-t FOCUSS and k-t SLR were calculated on CPU.

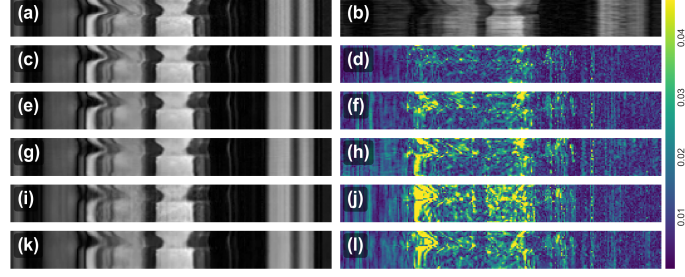


Figure 7.6: The comparison of reconstructions along temporal dimension with their error maps. (a) Ground Truth (b) Undersampled image by acceleration factor 9 (c,d) Proposed-B (e,f) 3D CNN (g,h) 3D CNN-S (i,j) k-t FOCUSS (k,l) k-t SLR

7.3.4 Variations of Architecture

In this section we show additional experiments to investigate the variants of the proposed architecture. First, we study the effects of recurrence over iteration and time, separately and jointly. In this study, we performed experiments on data set with undersampling factor 9, and the number of iterations was set to be 2 in order to simplify and speed up the training. Results are shown in Table 7.2, where we present the mean PSNR value via 3-fold cross validation. To isolate the effects of both recurrence in the module, we proposed to remove one of the recurrence each time. By removing the recurrence over time, the network architecture degrades to 4 CRNN-i + CNN layers, and it doesn't exploit temporal information in this case. If the recurrence over iterations is removed, the network architecture then becomes BCRNN-t + 4 CNN layers, without any hidden connections between iterations. Note that in all architectures, the last CNN layer only has 2 filters, which is used to simply aggregate the latent representation back to image space. Therefore, we employ a simple convolution layer for this. From Table 7.2, it can be observed that by removing any of the recurrent connections, the performance becomes worse compared with the proposed architecture with both recurrence jointly. This indicate that both of these recurrence contribute to the learning of the reconstruction. In particular, it has also been observed that by removing the temporal recurrence, the network's performance degrades greatly compared with the one removing the iteration recurrence. This can be explained that by removing the temporal recurrence, the problem degrades to a single frame reconstruction, while dynamic reconstruction has been proven to be much better than single frame reconstruction as there exists great temporal redundancies that can be exploited

Table 7.2: Performance comparisons on investigating the effects of each recurrence in the module. Reported results are the mean PSNR on data with undersampling factor 9 via 3-fold cross-validation. For this study, the number of iteration was set as 2.

Architectures	PSNR (dB)
4 CRNN-i + CNN (only iteration)	21.41
BCRNN-t + 4 CNN (only temporal)	26.62
BCRNN-t-i + 3 CRNN-i + CNN (Proposed)	27.98

Table 7.3: Performance comparisons with different model architectures. Reported results are the mean PSNR on data with undersampling factor 9 via 3-fold cross-validation. (FPT: forward pass time; BPT: backward pass time)

Architectures	PSNR (dB)	FPT	BPT	Training Time
4 BCRNN-t-i + CNN	34.18	0.94s	5.97s	96h
Proposed-A	33.28	0.45s	1.39s	38h
Proposed-B	34.38	0.90s	2.59s	58h

between frames.

In addition, experiments on some other variants of the architecture were performed, in particular, 4 layers of BCRNN-t-i with one layer of CNN, which has the highest capacity among all different combinations. Here we set the number of iterations to 10. It can be observed that incorporating temporal recurrent connections over all layers does improve the results over Proposed-A due to the more information propagated between frames. However, such design also increases the computations and more significantly, time required for training the network. Considering the trade off between performance and training time as well as the hardware constraints, we chose the particular design proposed. We agree that there could be more versions of the architectures that can lead to better performance and our particular design is by no means optimal. However, here we mainly aim to validate our proposed idea of exploiting both temporal and iterative reconstruction information for the problem, and thus the proposed architecture is sufficient to show this.

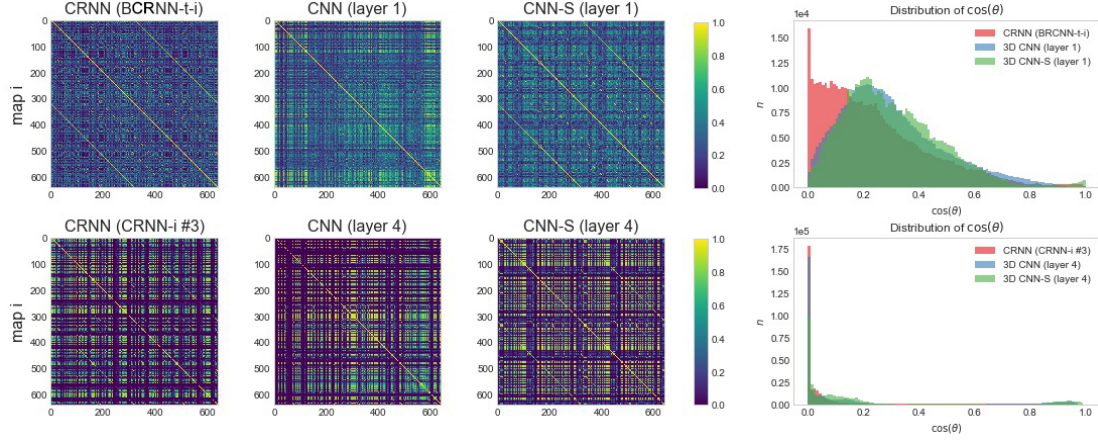


Figure 7.7: Cosine distances for the feature maps extracted from i th-layer of the subnetworks across 10 cascades/iterations. Top row shows $i = 1$, which corresponds to BCRNN-t-i unit for CRNN, 1st convolution layers for 3D-CNN and 3D-CNN-S. Bottom row shows $i = 4$, which corresponds to the third CRNN-i unit for CRNN, 4th convolution layers for 3D-CNN and 3D-CNN-S. In general, the distribution of $\cos(\theta)$ is closer to 0 for CRNN than for the CNN’s.

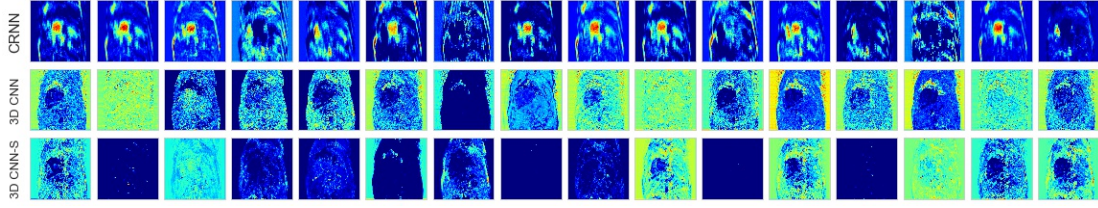


Figure 7.8: Examples of the feature maps from the CRNN-MRI (Proposed-A), 3D CNN and 3D CNN-S, at iteration 10.

7.3.5 Feature Map Analysis

In this section we further study whether the proposed architecture helps to obtain better feature representations. CRNN (Proposed-A), 3D-CNN and 3D-CNN-S all have the subnetworks composed of 5 units/layers with 64 channels for the first four, allowing us to directly compare the i -th layer of representations of the subnetworks for $i = 1, \dots, 4$. From one test subject, we extract the feature representations of the subnetwork across 10 cascades/iterations. By treating each channel as a separate feature map, we obtain 640 feature maps for each layer i aggregated across iterations. We use the cosine distance $d(A, B) = A^T B / \|A\| \|B\| = \cos(\theta)$ to compute the similarity between these activation maps for $i \in \{1, 4\}$. If two feature maps are orthogonal, then $\cos(\theta) = 0$ and if two feature maps are linearly correlated, then $\cos(\theta) = 1$. Geometrically, this supports the interpretation that if the cosine distance is small for all the

feature map pairs, then the network is likely to be capturing diverse patterns. The result is summarised in Fig. 7.7, where the similarity measure is visualised as a matrix, as well as their distributions is plotted for each network.

We can see that for both $i \in \{1, 4\}$, the layers from CRNN appears to have geometrically more orthogonal feature maps. One can also observe that in general, layer 1 has higher redundancy compared to layer 4. In particular, the diagonal yellow stripes can be observed for CNN-S and CRNN, due to parameter-sharing for each cascade. This is not observed in 3D-CNN, even though many features do have high similarity. In Fig. 7.8 we show examples of the feature maps from layer 4 (3rd CRNN-i for CRNN, 4th convolution layers for 3D-CNN and 3D-CNN-S) at iteration/cascade 10 of each network during the forward pass. We selected 16 feature maps out of 64 by firstly clustering them into 16 groups, and then randomly chose one feature map from each group to show as representative feature maps in Fig. 7.8. These feature maps show the activations learnt from different networks and are colour-coded (blue corresponds to low activation whereas red corresponds to high activation). We see that CRNN's features look significantly different from CNN. In particular, one can observe that some are activated by the dynamic region, and some are particularly sensitive to regions around the left and/or right ventricle.

7.4 Discussion

In this work, we have demonstrated that the presented network is capable of producing faithful image reconstructions from highly undersampled data, both in terms of various quantitative metrics as well as inspection of error maps. In contrast to unrolled deep network architectures proposed previously, we modelled the recurrent nature of the optimisation iteration using hidden representations with the ability to retain and propagate information across the optimisation steps. Compared with 3D CNN models, the proposed methods have a much lower network capacity but still have a higher accuracy, reflecting the effectiveness of our architecture. This is due to the ability of the proposed RNN units to increase the receptive field size while iteration

steps increase, as well as to efficiently propagate information across the temporal direction. In fact, for accelerated imaging, higher undersampling factors significantly add aliasing to the initial zero-filled reconstruction, making the reconstruction more challenging. This suggests that while the 3D CNN possesses higher modelling capacity owing to its large number of parameters, it may not necessarily be an ideal architecture to perform dynamic MR reconstruction, presumably because the simple CNN is not as efficient as propagating the information across the whole sequence. Besides, for the 3D CNN approaches, it is also observed that it is not able to denoise the background region. This could be explained by the fact that 3D CNN only exploits local information due to the small receptive field size it used, while in contrast, the proposed CRNN improves the denoising of the background region because of its larger receptive field sizes.

Furthermore, when exploring the intermediate feature activations, we observed that the pair-wise cosine distances for CRNN were smaller than those for the 3D CNNs. We speculate that this is because CRNN has hidden connections across the iterations allowing it to propagate information better and make the end-to-end reconstruction process more dynamic, generating less redundant representations. On a contrary, 3D CNNs need to rebuild the feature maps at every iteration, which is likely to increase repetitive computations. In addition, qualitatively, the activation map of CRNN showed high sensitivity to anatomical regions/dynamic regions. This is likely due to the fact that CRNN has increased receptive field size as well as temporal units, allowing the network to recognise larger/dynamic objects better. In CNNs, one can also observe that there are features activated by the myocardial regions, however, the activation is more homogeneous across the image, due to smaller receptive field size. This hints that CRNN can better capture high level information.

7.5 Conclusions

Inspired by variable splitting and alternate minimisation strategies, we have presented an end-to-end deep learning solution, CRNN-MRI, for accelerated dynamic MRI reconstruction, with

a forward, CRNN block implicitly learning iterative denoising interleaved by data consistency layers to enforce data fidelity. In particular, the CRNN architecture is composed of the proposed novel variants of convolutional recurrent unit which evolves over two dimensions: time and iterations. The proposed network is able to learn both the temporal dependency and the iterative reconstruction process effectively, and outperformed the other competing methods in terms of both reconstruction accuracy and speed for different undersampling rates.

Chapter 8

Dynamic MR Image Reconstruction Exploiting k - t Correlations

This chapter is based on:

- **C. Qin**, J. Schlemper, J. Duan, A. Price, J. Hajnal and D. Rueckert. k - t NEXT: Dynamic MR Image Reconstruction Exploiting Spatio-temporal Correlations. MICCAI, 2019.

8.1 Introduction

Dynamic MRI is a non-invasive imaging technique to monitor dynamic processes such as cardiac motion by acquiring data in a k - t space that contains both temporal and spatial information. It is well known that in dynamic MRI there exists significant correlations in k -space and time. In order to accelerate the acquisition, it is a common practice to acquire part of the desired k - t measurements and then reconstruct the images by exploiting spatio-temporal redundancies within the data.

Inspired by traditional k - t methods from the area of compressed sensing [128, 129, 244] for accelerated dynamic MR imaging, here we propose a novel dynamic MR image reconstruction Network with X- f Transform, termed k - t NEXT, which exploits the signal redundancies

in both x - f domain and image domain. In particular, the proposed k - t NEXT formulates the reconstruction process in an iterative fashion, where in each iteration, it consists of two sub-modules: a xf -CNN that learns to recover the true signals from aliased signals in x - f domain, and a convolutional recurrent neural network (CRNN) that exploits spatio-temporal redundancies in image domain. The dynamic reconstruction process thus alternates between x - f space and image space, which potentially enables the network to learn complementary features simultaneously from both domains. Experiments were performed on highly undersampled short-axis cardiac cine MR scans, where we show that the proposed model outperforms the current state-of-the-art dynamic MR reconstruction methods.

8.2 k - t NEXT for Dynamic MRI Reconstruction

Motivated by k - t BLAST [244] and k - t FOCUSS [129], we propose a dynamic image reconstruction Network with X- f Transform (k - t NEXT) to exploit the spatio-temporal correlations from both x - f space and image space. Specifically, as discussed and shown in Section 6.2, the solution of k - t FOCUSS can be expressed as the form as in Eq. 6.5. It consists of a baseline signal $\bar{\rho}$ and its residual encoding for the n -th estimate of the x - f signal $\rho^{(n)}$. By omitting the mathematical form of FOCUSS algorithm, we simplify the formulation as:

$$\rho^{(n)} = \bar{\rho} + \text{FOCUSS}(\rho^{(n-1)} - \bar{\rho}, \rho^{(n-1)}), \quad (8.1)$$

where FOCUSS here stands for the update function that receives input of a signal from previous iteration $\rho^{(n-1)}$ and its difference with baseline signal, i.e., $\rho^{(n-1)} - \bar{\rho}$.

Inspired by the k - t FOCUSS formulation, the proposed k - t NEXT formulates the iterative reconstruction process in an unfolded cascading way, as it has been shown to be a powerful technique in MR reconstruction [197, 215]. In each iteration, our proposed approach learns to reconstruct the true images by alternating between x - f and image spaces, so that the spatio-temporal redundancies can be jointly exploited from these two complementary domains. In particular, a xf -CNN is proposed for the recovery of signals in x - f domain inspired by the

traditional k - t method, and a variation of the CRNN-MRI network as proposed in Chapter 7 is adopted for the subsequent image space reconstruction. We can compactly represent a single iteration of the k - t NEXT as follows:

$$\rho^{(n)} = \text{DC}(\bar{\rho}_{rec}^{(n-1)}) + xf\text{-CNN}(\rho_{rec}^{(n-1)} - \bar{\rho}_{rec}^{(n-1)}), \quad (8.2a)$$

$$\sigma_{rec}^{(n)} = \text{CRNN}(\mathcal{F}_f \rho^{(n)}; \mathbf{v}^{(0)}), \quad \rho_{rec}^{(n)} = \mathcal{F}_f^H \sigma_{rec}^{(n)}, \quad (8.2b)$$

where $\sigma_{rec}^{(n)} \in \mathbb{C}^D$ denotes the complex-valued reconstructed image sequence at iteration n , and $\sigma_{rec}^{(0)} = \sigma_u$ is the acquired zero-filled undersampled images. Here $D = D_x D_y T$, in which D_x and D_y are width and height of the frame and T is the number of frames. $\mathcal{F}_f$ denotes the Fourier transform along f dimension, and $\rho_{rec}^{(n)}$ is the x - f spectral signal transformed from $\sigma_{rec}^{(n)}$, while $\rho^{(n)}$ stands for the intermediate reconstructed signal from xf -CNN. Also $\bar{\rho}_{rec}^{(n-1)}$ denotes the temporally averaged x - f signal (see Eq. (8.3)), DC stands for the data consistency layer [215], and $\mathbf{v}^{(0)} \in \mathbb{C}^M$ ($M \ll D$) is the acquired raw data. An illustrative diagram of k - t NEXT is shown in Fig. 8.1. We will introduce it in the following.

8.2.1 xf -CNN in x - f domain

Following the formulation in Eq. 6.5, here we propose to formulate the xf -CNN reconstruction as Eq. 8.2a, where instead of using model-based [244] or compressed sensing [129] algorithms to recover the true signals, we employ a stack of CNN layers to estimate the missing data based on other available points, typically within its vicinity in x - f space. In particular, here the x - f baseline signal $\bar{\rho}_{rec}^{(n)}$ is a temporal average of a sequence, i.e.,

$$\bar{\rho}_{rec}^{(n)} = \mathcal{F}^H \left[\sum_t \mathbf{v}^{(n)} ./ \max(1, \sum_t \delta(\mathbf{v}^{(n)})) \right]_{\text{tile} \rightarrow t}, \quad \delta(a) = \begin{cases} 0 & a = 0 \\ 1 & a \neq 0 \end{cases} \quad (8.3)$$

in which $\mathbf{v}^{(n)}$ is the k -space data that is Fourier transformed from $\sigma_{rec}^{(n)}$, the $./$ and $\max$ operation is performed element-wise, and $[\cdot]_{\text{tile} \rightarrow t}$ indicates the tile operation along temporal dimension. Thereby, xf -CNN learns to reconstruct residuals of each frame, which further exploits the signal

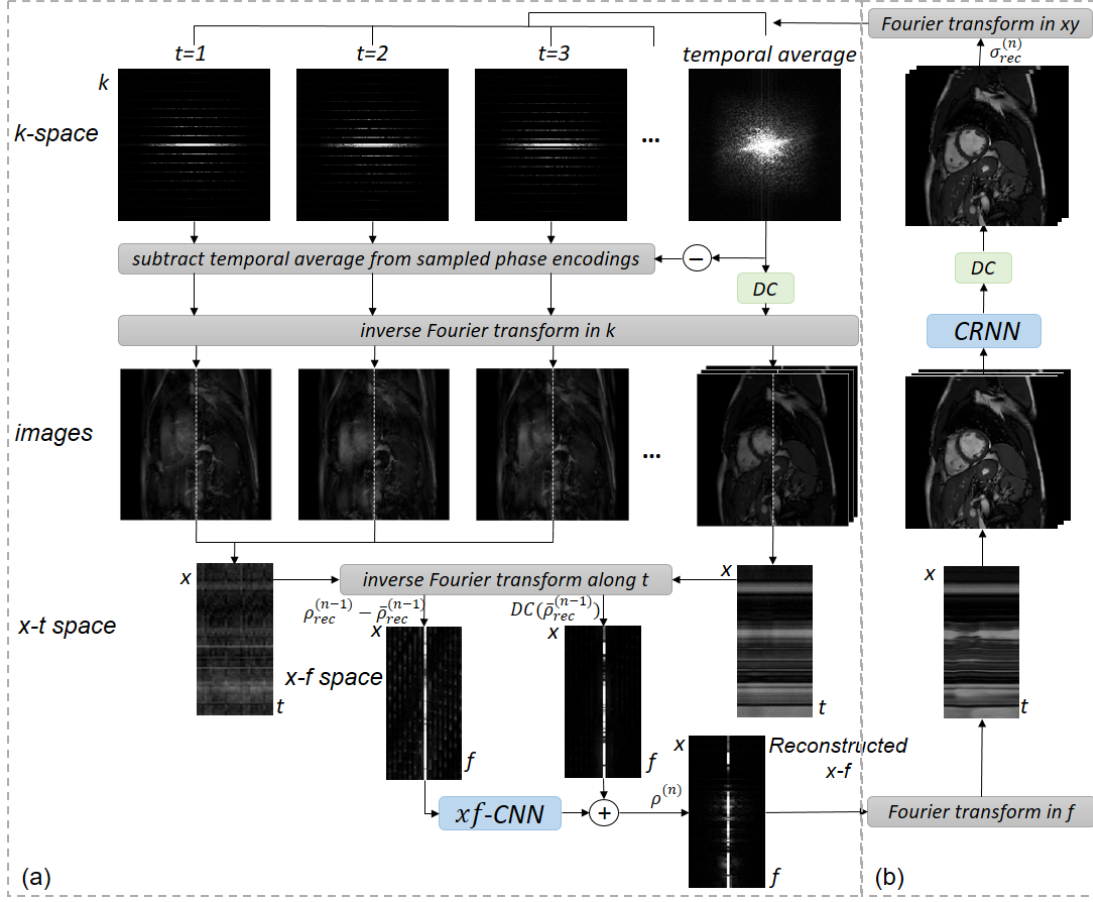


Figure 8.1: The k - t NEXT reconstruction diagram. True signals can be recovered by iteratively updating the reconstruction in both (a) x - f and (b) image domains via learning the xf -CNN and CRNN jointly. For mathematical notations, please refer to Eq. 8.2.

sparsity.

The illustrative diagram of x - f reconstruction is shown in Fig. 8.1(a). Specifically, we formulate the k - t to x - f transformation process as a **x - f transform layer** in the network. In details, the x - f transform layer receives input from k - t space data. For iteration n , the acquired k -space data is firstly averaged along t to yield a temporal average (Eq. 8.3), which is then subtracted from data at each time frame. To ensure data fidelity for the baseline estimate, here we propose to incorporate a data consistency (DC) term for $\bar{\rho}_{rec}^{(n-1)}$ at each frame separately. Then the subtracted data and temporally averaged data are inverse Fourier transformed to image space to obtain a sequence of aliased images and a data-consistent temporally averaged sequence. Each frequency-encoding position is then processed separately hereafter. The

image columns from aliased images or baseline images are then gathered and inverse Fourier transformed along t to yield an x - f image, corresponding to $\rho_{rec}^{(n-1)} - \bar{\rho}_{rec}^{(n-1)}$ and $\text{DC}(\bar{\rho}_{rec}^{(n-1)})$ respectively, which are then fed as inputs to xf -CNN for x - f space reconstruction (Eq. 8.2a). After the signal de-aliasing in x - f domain, another Fourier transform along f is adopted to transform the estimated x - f signal $\rho^{(n)}$ back to dynamic image space for the subsequent image space reconstruction (Eq. 8.2b).

8.2.2 *k-t* NEXT in complementary domains

Previous approaches [78] have shown that exploring cross-domain knowledge is beneficial for MR reconstruction task. Inspired by this, with the aim of exploiting redundancies in complementary domains, here we propose to learn a dynamic MR reconstruction network in both x - f and image spaces jointly. In particular, we employ the CRNN model for image space reconstruction due to its effectiveness in exploiting temporal redundancies with a relatively smaller network capacity [197]. Thus, in each cascade, the proposed *k-t* NEXT consists of a xf -CNN and a CRNN block, where it employs all 2D convolutions across spatial and temporal dimensions, in contrast to 3D convolutions used in the baseline method [215]. This enables the network to be more efficient and effective in learning useful and complementary features in x - f , spatial and temporal space simultaneously.

Given the training data S with undersampled data as input and fully sampled data as target, i.e., (σ_u, σ_t) in image space and (ρ_u, ρ_t) in x - f space, the network is trained end-to-end by minimising the pixel-wise mean squared error (MSE) between the reconstructed data and the ground truth fully sampled data:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n_S} \sum \left(\|\sigma_t - \sigma_{rec}^{(N)}\|_2^2 + \|\rho_t - \rho^{(N)}\|_2^2 \right), \quad (8.4)$$

where $\sigma_{rec}^{(N)}$ and $\rho^{(N)}$ denote the predicted image and x - f array at iteration N , i.e., the final output in each domain, $\boldsymbol{\theta}$ is the set of network parameters, and n_S is the number of training samples.

8.3 Experiments and Results

8.3.1 Dataset and Implementation Details

The dataset used in our experiments consists of 10 fully sampled complex-valued short-axis cardiac cine MRI. Each scan contains a single slice SSFP acquisition with 30 temporal frames. The raw data has 32-channel data with sampling matrix 192×190 , which was zero-filled to 256×256 , and the raw multi-coil data was then reconstructed to produce a single complex-valued image. In experiments, images were transformed back to k -space to simulate a fully sampled single-coil acquisition. A shear grid k - t Cartesian sampling pattern with four central lines (see Fig. 8.4(b)) was employed to undersample the k -space data to generate the undersampled input image sequences. The undersampling rate mentioned is stated with respect to the matrix size of the data, which is 192×190 .

A detailed network architecture of xf -CNN and CRNN for each iteration in the proposed k - t NEXT is shown in Fig. 8.2. Here xf -CNN is composed of 5 layers of 2D CNN with a residual connection from the baseline estimate. For the CRNN model, a variation of architecture [197] is employed which consists of 4 layers of bidirectional CRNN, 1 layer of 2D CNN, a residual connection and a DC layer. We used dilated convolutions with kernel size 3×3 and dilation factor (3, 3), and the number of cascade N was set to 4 for all comparison methods. The number of filters for each convolution was set to be 64. The network was implemented in PyTorch. During training, ADAM optimiser was employed with a learning rate of 10^{-4} . Data augmentation was performed on-the-fly, with random rotation, scaling, and elastic transformation. All evaluations were done via a 3-fold cross validation.

8.3.2 Results

In experiments, we compared our proposed approach (k - t NEXT) with different dynamic MR reconstruction methods, including compressed sensing method k - t FOCUSS [129], deep learning method CRNN-MRI [197], and DS+3DCNN [215] that incorporates data sharing (DS). To

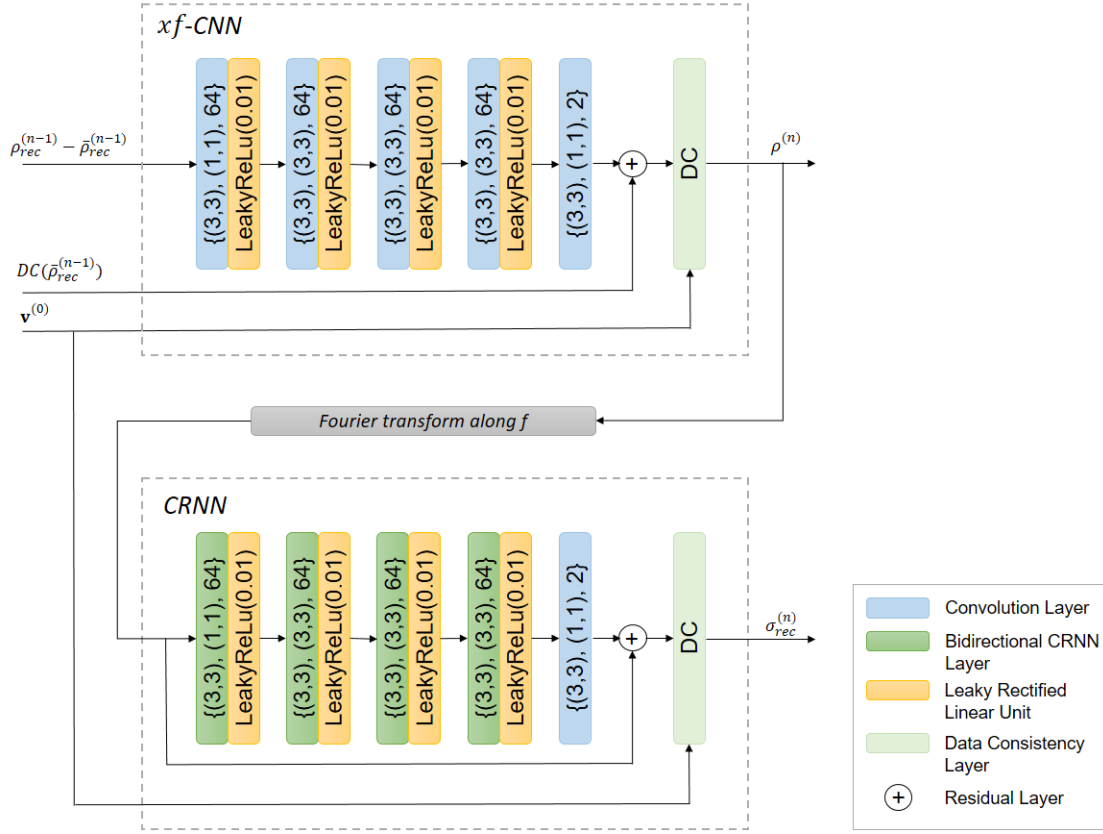


Figure 8.2: Detailed network architecture of xf -CNN and CRNN in k - t NEXT. Numbers inside CNN and BCRNN layer are kernel size, dilation factor, numbers of filters, and the negative slope for LeakyReLU is 0.01. DC stands for data consistency.

investigate the effectiveness of xf -CNN, an additional baseline approach is proposed which replaces all x - f reconstruction in k - t NEXT with DS component, termed DS+CRNN. In DS methods, we set the number of neighbouring frame as $n_{adj} \in \{0, 1, \dots, 5\}$ as in [215]. Note that for a fair comparison with our k - t NEXT, we modified the baseline approaches DS+3DCNN and DS+CRNN to learn the residual of a temporally averaged frame as well. Quantitative comparison results of different methods on dynamic cardiac data with undersampling rates 9 and 12 are presented in Table 8.1, where it compares the network capacity per cascade, peak-to-noise-ratio (PSNR), structural similarity index (SSIM) and high frequency error norm (HFEN) [197]. It can be seen that our proposed k - t NEXT can outperform other baseline methods by a large margin in terms of all these measures at different undersampling rates, with roughly the same level of network capacity. In particular, k - t NEXT performs better than its corresponding

Table 8.1: Comparison results of different methods on dynamic cardiac cine MRI with high undersampling rate 9 and 12. Best results are indicated in bold.

Method	k - t FOCUSS	CRNN-MRI	DS+3DCNN	DS+CRNN	k - t NEXT	
capacity	-	260,866	352,770	265,474	374,020	
9 $\times$	PSNR	29.52 (1.58)	32.45 (1.33)	33.47 (1.41)	33.24 (1.38)	34.23 (1.44)
	SSIM	0.951 (0.013)	0.969 (0.008)	0.975 (0.006)	0.975 (0.006)	0.979 (0.005)
	HFEN	0.340 (0.033)	0.249 (0.032)	0.214 (0.026)	0.215 (0.027)	0.196 (0.030)
12 $\times$	PSNR	28.14 (1.56)	31.30 (1.32)	32.46 (1.36)	32.34 (1.35)	33.18 (1.40)
	SSIM	0.937 (0.016)	0.962 (0.009)	0.969 (0.007)	0.970 (0.007)	0.975 (0.005)
	HFEN	0.382 (0.035)	0.282 (0.034)	0.242 (0.027)	0.239 (0.029)	0.225 (0.031)

DS pair, which indicates the merits of exploiting correlations in x - f space and complementary domains.

Additionally, we compared the qualitative results on 9 $\times$ undersampled data in Fig. 8.3, where it shows the reconstructed images along both spatial and temporal dimensions, as well as their corresponding error maps. It can be observed that our proposed model can faithfully recover the images with smaller errors especially around dynamic regions compared with other baseline methods. In particular, k - t NEXT produced visually sharper images than DS methods. This is reflected by the fact that, in contrast to DS approaches which fill in k -space data from neighboring frames and therefore could possibly generate averaged and smooth images, k - t NEXT directly estimates the missing data in x - f space. A visualisation of x - f reconstruction is also presented in Fig. 8.4, where it displays the reconstructed x - f image and its error map in comparison to the input aliased data. It can be observed that the aliasing artefacts were largely removed and the undersampled data were recovered to approximate the ground truth signals.

8.4 Conclusions

In this chapter, a novel deep learning based method, k - t NEXT (k - t NEtwork with X- f Transform), have been presented for highly undersampled dynamic MR image reconstruction. x - f -CNN is proposed to exploit correlations in k - t space via reconstructing the true signals from

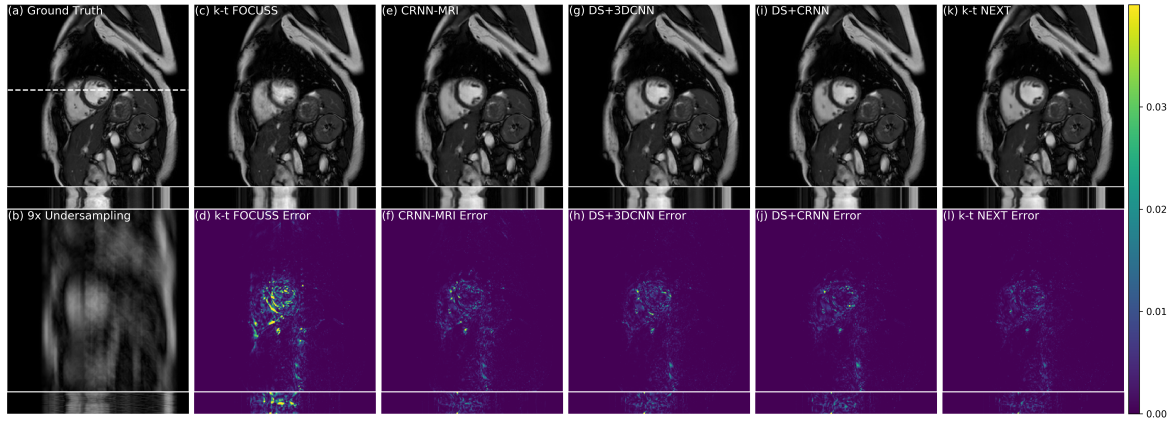


Figure 8.3: Comparison results on spatial and temporal dimensions with their error maps. A dynamic video is shown in supplementary materials for better visualisation.



Figure 8.4: Visualisation in x - f domain. (a) Ground Truth (b) k - t sampling pattern (c) $9\times$ undersampled data (d) Reconstructed x - f image (e) Error between (c) and (d).

aliased signals in x - f domain. Based on that, k - t NEXT is then proposed to learn to iteratively recover the images by alternating between the complementary x - f and image domains, where networks from both domains were trained jointly. Experimental results have shown that the proposed k - t NEXT outperforms state-of-the-art dynamic MR reconstruction methods in terms of both quantitative and qualitative performance.

Part III

Image Registration and Motion Estimation

Chapter 9

Background: Image Registration

9.1 Introduction

Image registration is the process of aligning multiple images so that their spatial correspondences can be established. In medical imaging, it is fundamental to relate spatial information of anatomical structures within the patient, together with spatial information about function and any pathology. It is an important precursor to an accurate and detailed medical image analysis and interpretation.

Image registration plays a key role in many medical image analysis applications, such as multimodal fusion, longitudinal studies, cross-sectional studies, etc. In clinical scenarios, different medical image modalities such as MRI, CT or Positron Emission Tomography (PET) show unique tissue features and are often acquired and fused for diagnostic and interventional purposes. As images from different modalities are often acquired with different scanners, this generally requires the alignment of corresponding features seen on different image modalities, providing the combination of complementary information. In addition, in longitudinal studies, it aims at the imaging of individuals at different points in time. Registering images from different time points enables the monitoring and quantification of anatomical changes in size or intensity over time. Besides, registering a large group of subjects to an atlas also allows the systematic quantitative analysis of the shape variability and morphometry across a population.

With the ever increasing amount of image data acquired, it is more and more desirable to register relevant images to extract clinical useful information for accurate comparisons within the same modality or the combination of complementary information between different modalities.

However, image registration still remains a challenging problem. In multi-modal registration, different modalities show different anatomical or functional properties of the patient being imaged, and the relationship between the intensity distributions of images is always unknown and complicated. This often makes the naive intensity based registration unfeasible. Also, there could be presence of features in one modality but missing in another, and thus such ambiguous correspondences can lead to image registration methods performing unexpectedly in those areas. Besides these, medical image acquisition artefacts such as motion artefacts, image noises or intensity inhomogeneities are challenging issues to registration problem, and robustness to these artefacts should also be taken into account when designing image registration techniques.

In this chapter, we will mainly focus on the non-rigid registration problem which models the deformable anatomical variabilities across individuals or tissue changes over time. We will first give an overview of its concepts and algorithms, and then present some applications and related work on both mono-modal registration and multi-modal registration.

9.2 Non-rigid Registration

In general, registration techniques can be decomposed into three components: a transformation that relates source images with the target images, a similarity metric which measures the similarity between source and target images, and an optimisation strategy to determine the transformation parameters by maximising the similarity function. The difference between rigid and non-rigid transformation mainly lies in their transformation. Non-rigid registration deforms images in a non-linear fashion, which are determined by a greater number of degrees or parameters than rigid or affine transformations. Here we will only discuss about non-rigid registration, also known as deformable registration, which is most related to our work.

9.2.1 Similarity Measures

Similarity metrics are defined to measure the similarities and correspondences between a pair of images, so that a transformation can be optimised for a registration. Let I_s denote the source image, and I_t denote the target image. Given a transformation T , the transformed source image is represented as I_s^T . Here we introduce some of the mostly used similarity measures for both mono-modal and multi-modal registration.

Sum of squared differences

One of the simplest and widely used voxel similarity measures is the sum of squared differences (SSD) between image intensities. It is defined as

$$SSD(I_t, I_s^T) = \frac{1}{\|\Omega\|} \sum_{x \in \Omega} (I_t(x) - I_s^T(x))^2, \quad (9.1)$$

where x is the voxel location, and Ω is the overlap image domain between target and source images. Note that this metric is optimal when two images only differ by Gaussian noise, and thus it is not applicable for intermodality registration.

Normalised cross-correlation

Cross-correlation assumes a linear intensity relationship between images when well aligned, which has a less strict assumption than SSD measure. Normalised cross-correlation (NCC) [213] is a general measure of statistical agreement, and it is defined as

$$NCC(I_t, I_s^T) = \frac{\sum_{x \in \Omega} (I_t(x) - \bar{I}_t) (I_s^T(x) - \bar{I}_s^T)}{\sqrt{\sum_{x \in \Omega} (I_t(x) - \bar{I}_t)^2} \sqrt{\sum_{x \in \Omega} (I_s^T(x) - \bar{I}_s^T)^2}}, \quad (9.2)$$

where $\bar{I}_t$ and $\bar{I}_s^T$ stand for the mean voxel intensity value in image I_t and I_s^T within domain Ω respectively. This NCC measure is invariant to linear intensity scaling.

Mutual information

For multi-modal image registration where linear intensity relationships cannot be assumed, information theory based similarity metrics can be defined to maximise the amount of shared information in two images. In particular, a 2D histogram $\mathcal{H}$ is constructed, where the matrix value $h(t, s)$ represents the occurrence of binned intensity pair (from $I_t(x)$ and $I_s^T(x)$) at the same voxel location. In details, the joint probability $p(t, s)$ of the intensity pair (t, s) from $I_t(x)$ and $I_s^T(x)$ can be represented as $p(t, s) = h(t, s)/N$, where N is the number of samples in the joint histogram. Mutual information (MI) takes into account the joint entropy, but also the marginal entropies for each image. The marginal probabilities of the occurrence of t in image I_t is estimated as $p(t) = \sum_s p(t, s)$, and similarly the marginal probabilities for image I_s^T is $p(s) = \sum_t p(t, s)$. By using the Shannon formula [221] for entropy, the marginal entropy is formulated as:

$$\begin{aligned} H(I_t) &= - \sum_t p(t) \log(p(t)), \\ H(I_s^T) &= - \sum_s p(s) \log(p(s)), \end{aligned} \tag{9.3}$$

and the joint entropy $H(I_t, I_s^T)$ is given by:

$$H(I_t, I_s^T) = - \sum_t \sum_s p(t, s) \log(p(t, s)). \tag{9.4}$$

Mutual information measures how well one image explains the other, and it is defined as:

$$MI(I_t, I_s^T) = H(I_t) + H(I_s^T) - H(I_t, I_s^T). \tag{9.5}$$

By maximising mutual information [164, 256], a transformation that aligns the source image and target image could be found.

Normalised mutual information

In addition to mutual information, alternative normalisations of joint entropy have been proposed. Studholme et al. [233] has proposed normalised mutual information (NMI) which has

been shown to be more robust to changes in overlap regions in two images during registration process. NMI is defined as:

$$NMI(I_t, I_s^T) = \frac{H(I_t) + H(I_s^T)}{H(I_t, I_s^T)}. \quad (9.6)$$

9.2.2 Techniques

Over the years, there have been a large amount of research going on in the area of image registration. Here we describe some of the representative algorithms for deformable image registration, including both traditional methods and deep learning based approaches. We will mainly focus on methods that will be used in the following chapters. For more detailed survey, please refer to [231, 284].

Free form deformation

Free form deformation (FFD) is an efficient transformation model that is manipulated by an underlying mesh of control points. In contrast to thin plate splines which allow arbitrary control points configurations, FFDs are parameterised by a regular mesh of uniformly spacing control points. FFDs based on B-splines have been widely used for 3D deformable image registration [209].

Assume control points are uniformly spaced on a $n_x \times n_y \times n_z$ mesh, with uniform space $\delta_x, \delta_y, \delta_z$ in each axis. Displacement vector $\phi_{i,j,k}$ represents the control points at location $\mathbf{x} = (i, j, k)$. In FFD, the displacement at location $\mathbf{x} = (x, y, z)$ can be expressed as the form:

$$T(x, y, z) = \sum_{l=0}^3 \sum_{m=0}^3 \sum_{n=0}^3 \theta_l(u) \theta_m(v) \theta_n(w) \phi_{i+l, j+m, k+n}, \quad (9.7)$$

where $i = \lfloor \frac{x}{\delta_x} \rfloor - 1$, $j = \lfloor \frac{y}{\delta_y} \rfloor - 1$, $k = \lfloor \frac{z}{\delta_z} \rfloor - 1$, $u = \frac{x}{\delta_x} - \lfloor \frac{x}{\delta_x} \rfloor$, $v = \frac{y}{\delta_y} - \lfloor \frac{y}{\delta_y} \rfloor$, $w = \frac{z}{\delta_z} - \lfloor \frac{z}{\delta_z} \rfloor$,

and θ_l represents the l -th basis function of the B-splines:

$$\begin{aligned}
 \theta_0(s) &= (1 - s)^3/6 \\
 \theta_1(s) &= (3s^3 - 6s^2 + 4)/6 \\
 \theta_2(s) &= (-3s^3 + 3s^2 + 3s + 1)/6 \\
 \theta_3(s) &= s^3/6
 \end{aligned} \tag{9.8}$$

Since the basis functions for cubic B-splines have a limited support, displacement at a control point only affects the transformations in its neighborhood, which makes the computation efficient even for a large number of control points.

A hierarchical, coarse-to-fine strategy is often used when performing non-rigid registration using FFDs [209, 219], where FFD parameters are optimised on multi-resolution levels. The idea is to first use a FFD with a relatively large control point spacing to capture the global deformations, followed by a control point mesh sub-division with half the control point spacing to capture more localised deformations. This process can be repeated for a number of resolution levels as specified. A pyramid of blurred and downsampled images corresponding to the resolution level can be used to facilitate the registration optimisation. In addition, when source image and target image are not in the same coordinate space, a combination of rigid or affine transformation with a subsequent non-rigid transformation model is usually carried out.

Demons algorithm

Demons algorithm is widely used for non-rigid 3D medical image registration. Thirion [239] proposed to consider non-rigid registration problem as a diffusion process, which is based on a concept of demons that is introduced by Maxwell to solve the Gibbs paradox in thermodynamics. The main idea is to consider the source image as a deformable grid, whose vertices are particles which can be classified as ‘inside’ or ‘outside’ particles. A contour of object in the target image is considered as a membrane where demons are scattered along. The registration process is to diffuse the grid through the contour, where those demons act as effectors, locally pushing the grid inside the contour if the corresponding point of the grid is labelled as ‘inside’

and outside the contour if labelled as 'outside'. Three different non-rigid matching algorithms were proposed using the concept of diffusing models [239], one using all intensity levels, one using only contour points, and one based on already segmented images. It has been shown that this algorithm can be considered as an approximation of a second order gradient descent on the sum of square of intensity differences criterion [189].

Based on demons algorithm, Vercauteren et al. [254] proposed a diffeomorphic demons method, in which it adapted the optimisation procedure underlying the demons algorithm to a space of diffeomorphic transformations. The solution is computationally efficient and produces accurate and smooth transformations.

Optical flow estimation

Optical flow is the pattern of relative motion of object between successive frames of a temporal sequence. Optical flow techniques have also been used for image registration task, with the assumption that image intensities of a particular point do not change between frames. This assumption can be formulated as:

$$I(x, y, z, t) = I(x + \Delta x, y + \Delta y, z + \Delta z, t + \Delta t). \quad (9.9)$$

Assuming the movement between frames to be small, by using a Taylor expansion of Eq. 9.9 and ignoring higher-order terms, it can be derived as:

$$\frac{\delta I}{\delta x} \frac{\Delta x}{\Delta t} + \frac{\delta I}{\delta y} \frac{\Delta y}{\Delta t} + \frac{\delta I}{\delta z} \frac{\Delta z}{\Delta t} + \frac{\delta I}{\delta t} = 0. \quad (9.10)$$

This can also be formulated as:

$$\frac{\delta I}{\delta t} + \nabla I \cdot \mathbf{u} = 0, \quad (9.11)$$

where $\frac{\delta I}{\delta t}$ represents the temporal difference between images, ∇I is the spatial gradient of the image, and $\mathbf{u}$ corresponds to the motion between the two images. Additional smoothness constraints on displacement field $\mathbf{u}$ can be added to obtain a reliable estimate of the flow field.

Velocity field based registration

To take into account the group structure of diffeomorphism, many works have been proposed to parameterise dense transformation fields with time-varying velocity fields [29, 242] or stationary velocity field [16], which generate geometrical deformations via the integration of an Ordinary Differential Equation (ODE). These include work such as large deformation diffeomorphic metric mapping (LDDMM) [29], DARTEL algorithm [17], Hernandez et al. [109] etc.

Specifically, for time-varying velocity fields, the transformation $\mathbf{T}(\mathbf{x}) = \mathbf{y}$ is modelled as a time-dependent flow $w(\mathbf{x}, t)$, where $\mathbf{w}(\mathbf{x}, 0) = \mathbf{x}$ and $\mathbf{w}(\mathbf{x}, 1) = \mathbf{y}$. The flow is defined via the ordinary differential equation:

$$\frac{\delta \mathbf{w}(\mathbf{x}, t)}{\delta t} = \mathbf{v}(\mathbf{w}(\mathbf{x}, t), t). \quad (9.12)$$

By integrating the time-varying velocity field $\mathbf{v}(\mathbf{w}(\mathbf{x}, t), t)$, the displacement field can be obtained as:

$$\mathbf{T}(\mathbf{x}) = \int_0^1 \mathbf{v}(\mathbf{w}(\mathbf{x}, t), t) dt. \quad (9.13)$$

Using time-varying velocity fields can be memory consuming and computationally expensive. Alternatively, diffeomorphism can be parameterised with stationary velocity fields, where velocity fields do not depend on time, i.e.,

$$\frac{\delta \mathbf{w}(\mathbf{x}, t)}{\delta t} = \mathbf{v}(\mathbf{w}(\mathbf{x}, t)). \quad (9.14)$$

To model deformations using stationary velocity fields, Arsigny et al. [16] proposed to define the exponential $\exp(V)$ of a smooth vector field $V(x)$ as the flow at time 1 of the stationary ODE $\dot{x} = V(x)$, based on Lie Group theory. In this setting, the diffeomorphic deformation field then becomes $\mathbf{w}(1) = \exp(v)$. To compute the exponentials, a fast and efficient algorithm called ‘scaling and squaring’ can be adopted, and it can be generalised to vector fields in the following way:

- **Scaling Step:** divide $V(x)$ by a factor of 2^N , so that $V(x)/2^N$ is close enough to zero.
- **Exponentiation Step:** $\mathbf{w}(1/2^N)$ is computed with a first-order explicit numerical scheme.
- **Squaring Step:** N recursive squaring of $\mathbf{w}(1/2^N)$ yield an estimation of $\mathbf{w}(1)$

Deep learning based registration

In recent years, many deep learning approaches have been proposed for image registration. They can be mainly divided into two groups: supervised, and unsupervised methods. The general framework is to estimate the deformation field between source image and target image, using deep convolutional neural networks (CNN) to learn useful features or model deformation priors. Given a test source image and target image, the trained models are able to predict the deformation fields directly. In supervised learning settings, they rely on ground truth deformation fields that are generated synthetically or obtained via classical registration methods [47, 141, 230]. However, ground truth deformations derived from conventional registration method can restrict the type of deformations that are learnt. In comparisons, unsupervised registration methods do not rely on generated deformation fields. Most current approaches use convolutional neural networks [23, 194] or probabilistic framework [58, 142] with a spatial transformation function [123] as introduced in Section 2.3.4, to warp the source image to the target one using the estimated deformation field. Then the networks are trained by minimising the conventional image similarity metrics between the target image and warped source image, such as SSD and NCC. Such unsupervised deep learning based registration methods showed competitive performance against those classical registration methods while at a faster registration speed. For a detailed survey, please refer to [105].

9.3 Related Work

In this section, some of the related work on deformable image registration will be reviewed, both on mono-modality registration and multi-modality registration.

9.3.1 Deformable Mono-modal Registration

Mono-modal registration has been widely studied. It is assumed that similar anatomical structures in source and target images share similar intensity values, as both volumes are imaged by the same imaging device. Thus criteria such as SSD and cross-correlations can be employed to align mono-modal images [231]. However, relying on intensity information only could lead to some ambiguous matching. Therefore, a number of researchers have proposed to introduce some local information that could represent the geometric structures to increase the dimensionality of feature space [185, 222]. Over the years, there has been extensive work on mono-modal image registration, and some of these representative works include free-form deformations using B-splines [209], elastic-type models [62], Demons [189, 239], and discrete methods [93]. Diffeomorphic registration methods such as diffeomorphic demons [254], LDDMM [29], and standard symmetric normalisation (SyN) [18] have also been proposed. These non-learning based approaches optimise within the space of displacement vector fields for each image pair, thus they are slow in terms of registration speed.

There are also some works recently that propose to use deep neural networks to learn the function for medical image registration [105]. One category of these approaches rely on ground truth deformation fields to learn the regression. For instance, Cao et al. [47] proposed to employ a CNN architecture to map input patches of a pair of 3D volume to their corresponding displacement vector, where the ground truth is generated using some conventional registration methods, while Uzunova et al. [245] proposed to synthesise the ground truth deformations using statistical appearance models (SAMs). For unsupervised deformable mono-modal image registration, most current approaches proposed to use convolutional neural networks [23, 65, 194] to estimate the deformation fields, together with a spatial transformation function [123] that serves as a bilinear sampler to transform the source image to the target one. To train the networks, image similarity metrics such as intensity differences or cross-correlations can be used as loss functions. Smoothness terms that penalise the gradients of deformation fields are often added to ensure the generated deformation fields to be smooth and regular. Instead of using specific similarity metrics, Fan et al. [80] proposed a similar registration network

which was trained along with learning a similarity measure by using a discriminator network to judge whether a pair of images are sufficiently aligned or not. In addition, some probabilistic models for diffeomorphic registration [58, 142] have also been proposed, where unlike other deep learning models which estimate deformation fields directly, they predict the stationary velocity fields, from which the deformation can be derived by an exponentiation layer. All these deep learning based methods present competitive performances against conventional methods, and they have a much faster registration speed compared to those non-learning based approaches.

Cardiac Motion Estimation

One of the applications of deformable mono-modal registration is to estimate the motion pattern of a sequence, such as cardiac motion estimation. Cardiac motion tracking can be viewed as a 4D registration problem. Traditional methods commonly extended classical optical flow or image registration methods for cardiac motion estimation [64, 223, 225, 240]. For instance, Perperidis et al. [190] proposed to use B-spline based FFDs for the spatio-temporal alignment of cardiac MR sequences, where the 4D B-spline was separated into spatial and temporal components. Myocardial incompressibility can be used in registration methods for cardiac motion estimation, for example, De Craene et al. [64] optimised a 4D velocity field parameterised by B-Spline spatio-temporal kernels to introduce temporal consistency with a regularisation term based on incompressibility of myocardial tissue. Besides, Shi et al. also proposed to [225] combine complementary information from both untagged and 3D tagged MR sequences to estimate myocardial motion using a series of FFDs [209], which showed a clear improvement in terms of accuracy over using either 3D tagged or untagged MR image information alone.

9.3.2 Deformable Multi-modal Registration

Multi-modal image registration is more challenging than mono-modal registration as the choice of their objective function is a harder task [231]. One category of the classical and standard methods for multi-modal registration are information theory based approaches, which utilise

mutual information (MI) as a similarity measure to align multi-modal images. It showed a great success in rigid registration of multi-modal medical images, and later its variations such as normalised MI and local MI etc. [139, 191, 256] have also been proposed to tackle deformable registration. However, such methods are often based on intensity probability distribution, and thus ignore spatial information of anatomical structures. An alternative way to address multi-modal image registration problem is to reduce the problem to a mono-modal one. They either synthesise one modality from another or map both modalities to a common domain. In order to reduce the appearance gap between different modalities, image synthesis can be achieved by taking advantage of prior knowledge on physical properties of imaging devices [205, 267] or capturing intensity relationships using learning-based methods [46]. As to mapping both modalities to a common space, the assumption is that both modalities share the same anatomical structure or feature, and thus can be used to establish meaningful correspondences. In principle, most approaches apply filters to extract geometrical information, such as edge information [171] and intensity gradient information [100], that can be used in mono-modal settings. For instance, a modality independent neighborhood descriptor (MIND) [108] was proposed based on the similarity of small image patches within one image, aiming to extract the distinctive structure in a local neighborhood that was preserved across modalities.

In recent years, some deep learning approaches have also been proposed to address the multi-modal image registration problem. In supervised setting, these methods require ground truth deformation fields during the training process. However, as ground truth deformations are rarely available, they commonly synthetically generate geometric deformations as ground truth and then transform one of the image pairs [236]. Some other methods employ kind of weakly-supervised way for image registration, where they rely on the alignment of multiple labelled corresponding anatomical structures for individual image pairs during the training [113, 114]. Instead of estimating deformation fields directly from neural networks, Tanner et al. [238] proposed to synthesise one modality to another using cycle-GAN, and then deformation fields were estimated between translated moving image and fixed image by employing conventional methods. Alternatively, Simonovsky et al. [227] used a CNN to explicitly estimate image similarity in the multi-modal case. Gradient descent was then used to iteratively update defor-

mation field related parameters given the learned similarity metrics. However, relevant works on fully unsupervised end-to-end deep learning approaches for multi-modal image registration are still limited, and it presents a promising direction for future research.

9.4 Conclusions

This chapter presents an overview of basic concepts and algorithms for deformable image registration. In particular, it introduces some commonly used similarity measures and techniques, which will be adopted in the work presented in this thesis. Related work on both mono-modal and multi-modal image registration as well as their applications have also been reviewed in this chapter, especially on deep learning based methods that have been widely researched in recent years. In the remaining chapters of this part, two novel deep learning models for cardiac motion estimation and multi-modal deformable image registration will be presented.

Chapter 10

Joint Motion Estimation and Segmentation of Cardiac MRI

This chapter is based on:

- **C. Qin**, W. Bai, J. Schlemper, S. Petersen, S. Piechnik, S. Neubauer and D. Rueckert. Joint Learning of Motion Estimation and Segmentation for Cardiac MR Image Sequences. MICCAI (2018)
- **C. Qin**, W. Bai, J. Schlemper, S. Petersen, S. Piechnik, S. Neubauer and D. Rueckert. Joint Motion Estimation and Segmentation from Undersampled Cardiac MR Image. International Workshop on Machine Learning for Medical Image Reconstruction, 2018: 55-63.

10.1 Introduction

Cardiac MRI is one of the reference methods to provide qualitative and quantitative information of the morphology and function of the heart, which can be utilised to assess cardiovascular diseases. Both cardiac MR image segmentation and motion estimation are crucial steps for the dynamic exploration of the cardiac function, which enable the accurate quantification of regional

function measures such as changes in ventricular volumes and the elasticity and contractility properties of the myocardium [223]. Traditionally, most approaches consider segmentation and motion estimation as two separate problems. Nevertheless, these two tasks are known to be closely related [52, 243], and learning meaningful representations for one problem should be helpful to learn representations for the other one. On the other hand, as we know, one limitation of the cardiovascular MR is the low acquisition speed. Nevertheless, in most cases, perfect reconstructions are not always necessary as long as the images allow to obtain accurate clinically relevant parameters. Therefore, instead of firstly recovering non-aliased images, it may be more effective to estimate the final results directly from undersampled MR data and also to make such estimations as accurate as possible.

This chapter will first introduce a joint deep learning network for predicting the segmentation and motion estimation simultaneously from fully sampled cardiac MR sequences. In particular, the proposed architecture consists of two branches: one is an unsupervised Siamese style spatial transformer network for cardiac motion estimation, which exploits multi-scale features and recurrent units to accurately predict sequences of motion fields while ensuring spatio-temporal smoothness; and the other one is a segmentation branch which takes advantage of the joint feature learning to enable weakly-supervised segmentation for temporally sparse annotated data. The problem is formulated as a composite loss function optimised by training both tasks simultaneously.

Then the chapter will present a work that extends the above joint framework for cardiac motion estimation and segmentation directly from undersampled cardiac MR data, bypassing the MR reconstruction process. The framework is formulated by incorporating supervision from fully sampled MR image pairs in addition to the composite loss function. Experiments indicate that results learned directly from undersampled data are reasonably accurate and are close to predictions from fully-sampled data. This could potentially lead to future works that enable fast and accurate analysis in an integrated MRI reconstruction and analysis pipeline.

10.2 Joint Learning of Motion Estimation and Segmentation

10.2.1 Methods

Our goal is to realise the simultaneous motion estimation and segmentation for cardiac MR image sequences. Here we construct a unified model consisting of two branches: an unsupervised motion estimation branch based on a Siamese style recurrent multi-scale spatial transformer network, and a segmentation branch based on a fully convolutional neural network, where the two branches share a joint feature encoder. The overall architecture of the model is shown in Fig. 10.1.

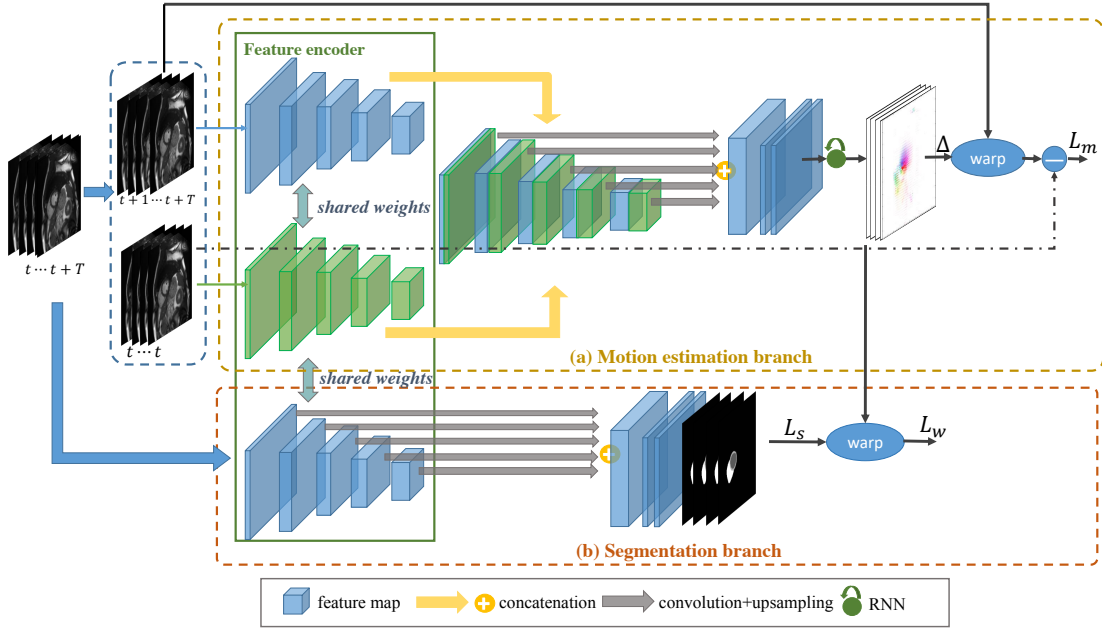


Figure 10.1: The overall schematic architecture of proposed network for joint estimation of cardiac motion and segmentation. (a) The proposed Siamese style multi-scale recurrent motion estimation branch. (b) The segmentation branch which shares the joint feature encoder with motion estimation branch. The architecture for feature encoder is adopted from VGG-16 net before FC layer. Both branches have the same head architecture as the one proposed in [21], and the concatenation layers of motion estimation branch are from last layers at different scales of the feature encoder.

Unsupervised Cardiac Motion Estimation

Deep learning methods normally rely heavily on the ground truth labelled data. However, in problems of cardiac motion estimation, dense transformation maps between frames are rarely available. Inspired by the success of spatial transformer network [43, 123, 187] which effectively encodes optical flow to describe motion, here we propose a novel Siamese style multi-scale recurrent network for estimating the cardiac motion of MR image sequences without supervision effort. A schematic illustration of the model is shown in Fig. 10.1(a).

The task is to find a sequence of consecutive optical flow representations between the target frame I_t and the source frames $I_{t+1}, I_{t+2}, \dots, I_{t+T}$, where the output is pixel-wise 2D motion fields Δ representing the displacement in x and y directions. In order to realise this, the proposed network mainly consists of four components: a Siamese network for the feature extraction of both target frame and source frame; a multi-scale concatenation of features from pairs of frames; a convolutional recurrent unit (RNN) which propagates information along temporal dimension; and a sampler that warps the source frame to the target one by using the estimated displacement field. In details, inspired by the success of cardiac segmentation network proposed in [21], we determine the Siamese feature encoder as the one in [21] which is adapted from VGG-16 net. For the combination of information from frame pairs, motivated by the traditional multi-level registration method [209], here we propose to concatenate multi-scale features from both streams of Siamese network to exploit information at different scales. This is followed by a convolution and upsampling operation back to the original resolution, and then they are combined using a concatenation layer. In addition, in order to exploit information from consecutive frames and also to ensure the spatio-temporal smoothness of the estimated motion fields, we additionally incorporate a convolutional simple RNN with tanh function at the last layer to propagate motion information along the temporal dimension and to estimate flow with two feature maps $\Delta = (\Delta x, \Delta y; \theta_\Delta)$ corresponding to displacements for the x and y dimensions, where the network is parameterised by θ_Δ . Finally, the source frames I_{t+k} are transformed using bilinear interpolation to the target frame, which can be expressed as $I'_{t+k}(x, y) = \Gamma\{I_{t+k}(x + \Delta_{t+k}x, y + \Delta_{t+k}y)\}$.

To train the spatial transformer, we optimise the network by minimising the pixel-wise mean squared error between the transformed frames and the target frame. To ensure local smoothness, we penalise the gradients of flow map by using an approximation of Huber loss proposed in [43], namely $\mathcal{H}(\delta_{x,y}\Delta_t) = \sqrt{\epsilon + \sum_{i=x,y}(\delta_i\Delta_i^2)}$ and similarly, we use a regularisation term $\mathcal{H}(\delta_t\Delta) = \sqrt{\epsilon + \sum_{i=x,y,t}\delta_i\Delta_i^2}$ to constrain the flow to behave smoothly in temporal dimension, where $\epsilon = 0.01$. Therefore, the loss function can be described as follows:

$$\mathcal{L}_m = \frac{1}{T} \sum_{k=1}^T [\|I_t - I'_{t+k}\|^2 + \alpha \mathcal{H}(\delta_{x,y}\Delta_{t+k})] + \beta \mathcal{H}(\delta_t\Delta), \quad (10.1)$$

where T is the number of sequence, α and β are regularisation parameters to trade off between image dissimilarity, local and temporal smoothness.

Joint Model for Cardiac Motion Estimation and Segmentation

As we know, motion estimation and segmentation tasks are closely related, and previous works in computer vision domain have shown that the learning of one task is able to benefit the other [52, 243]. Motivated by the success of self-supervised learning which learns features from intrinsic freely available signals [8, 71], here we propose a joint learning model for cardiac motion estimation and segmentation, where features learned from unsupervised (or self-supervised) motion estimation are exploited for segmentation. By coupling the motion estimation and segmentation network, the proposed approach can be viewed as a weakly-supervised method with temporally sparse annotated data while motion estimation facilitates the feature learning by exploring those unlabelled data. The schematic architecture of the unified model is shown in Fig. 10.1.

In details, the proposed joint model consists of two branches: the motion estimation branch, and the segmentation branch based on the effective network proposed in [21]. Here both branches share the joint feature encoder (Siamese style network) as shown in Fig. 10.1, so that the features learned can better capture the useful related representations for both tasks. Here a categorical cross-entropy loss $\mathcal{L}_s = -\sum_{l \in L} y_l \log(f(x_l; \Theta))$ on labelled data set L is used

for segmentation branch, in which we define x_l as the input data, y_l as the ground truth, and f is the segmentation function parameterised by Θ . In addition, to further exploit the input unlabelled data, we add an additional spatial transformer in segmentation branch, which warps the predicted segmentation to the target frame using the motion fields estimated from motion estimation branch. Similarly, a categorical cross-entropy loss $\mathcal{L}_w = -\sum_{n \in U} y_l \log(f_w(x_n; \Theta))$ is used between the warped segmentations and the target, where U stands for unlabelled data set, and f_w is f plus the warp operation. This component mainly works as a regularisation for the motion estimation branch, which is supposed to improve the estimation around boundaries.

As a result, a composite loss function consisting of image similarity error, smoothness penalty of motion fields, and pixel-wise cross entropy segmentation losses with the softmax function can be defined as follows:

$$\mathcal{L} = \mathcal{L}_m + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_w, \quad (10.2)$$

where λ_1 and λ_2 are trade-off parameters for different tasks. To initialise the joint model, we first train the motion estimation branch using all the available data we have. Then we fix the weights of the shared feature encoder, and train the segmentation branch with the available annotated data. Lastly, we jointly train both branches by minimising the composite loss function on training set.

10.2.2 Experiments and Results

Experiments were performed on 220 short-axis cardiac MR sequences from UK Biobank study. Each scan contains a sequence of 50 frames, where manual segmentations of left-ventricular (LV) cavity, the myocardium (Myo) and the right-ventricular (RV) cavity are available on ED and ES frames. A short-axis image stack typically consists of 10 image slices, and the pixel resolution is $1.8 \times 1.8 \times 10.0 \text{ mm}^3$. For pre-processing, all training images were cropped to the same size of 192×192 , and intensity was normalised to the range of $[0,1]$. In our experiments, we split the data into 100/100/20 for training/testing/validation. Parameters used in the loss

Table 10.1: Evaluation of segmentation accuracy for the proposed joint model and the baseline (Seg only) method in terms of the Dice Metric (mean and standard deviation).

Method	LV	Myo	RV
Seg only	0.9217 (0.0450)	0.8373 (0.0309)	0.8676 (0.0513)
Joint model	0.9348 (0.0408)	0.8640 (0.0295)	0.8861 (0.0453)

function were set to be $\alpha = 0.001$, $\beta = 0.0001$, $\lambda_1 = 0.01$ and $\lambda_2 = 0.001$, which were chosen via validation set. The number of image sequence for RNN during training was $T = 10$, and a learning rate of 0.0001 was used. Data augmentation was performed on-the-fly, with random rotation, translation, and scaling.

Evaluation was performed with respect to both segmentation and motion estimation. We first evaluated the segmentation performance of the joint model by comparing it with the baseline method, i.e., training the segmentation branch only (Seg only). Results reported in Table 10.1 are Dice scores computed with manual annotations on LV, Myo, and RV. It shows that the proposed joint model significantly outperforms the results of Seg only on all three structures with $p \ll 0.001$ using Wilcoxon signed rank test, especially on Myo where motion normally affects the segmentation accuracy greatly. This indicates the merits of joint feature learning, where features explored by motion estimation are beneficial for segmentation task.

We also evaluated the performance of motion estimation by comparing the results obtained using a B-spline free-form deformation (FFD) algorithm¹ [209], network for Motion only, and the joint model. We warped the segmentations of ES frame to ED frame by using the estimated motion fields, and mean contour distance (MCD) and Hausdorff distance (HD) were computed between the transformed segmentations and the segmentations of ED frame. Table 10.2 shows the comparison results of these methods. It can be observed that both of the proposed methods outperform FFD registration method in terms of MCD and HD on all the three structures ($p \ll 0.001$) and similarly, the joint model shows better performance than the model trained for motion estimation only ($p \ll 0.001$ on LV and RV, and $p < 0.01$ on Myo). Additionally, we compared the test time needed for motion estimation on 50 frames of a single slice in a cardiac cycle, and results indicated a faster speed of proposed methods compared to FFD.

¹<https://github.com/BioMedIA/MIRTK>

Table 10.2: Evaluation of motion estimation accuracy for FFD, proposed model for Motion only and the proposed joint model in terms of the mean contour distance (MCD) and Hausdorff distance (HD) in mm (mean and standard deviation). Time reported is testing time on 50 frames in a cardiac cycle per slice.

Method	MCD			HD			Time
	LV	Myo	RV	LV	Myo	RV	
FFD	1.83(0.53)	2.47(0.74)	3.53(1.25)	5.10(1.28)	6.47(1.69)	12.04(4.85)	13.22s
Motion only	1.55(0.49)	1.23(0.30)	3.14(1.12)	4.20(1.04)	3.51(0.88)	11.72(4.28)	1.23s
Joint model	1.30(0.34)	1.19(0.26)	3.03(1.08)	3.52(0.82)	3.43(0.87)	11.38(4.34)	2.80s

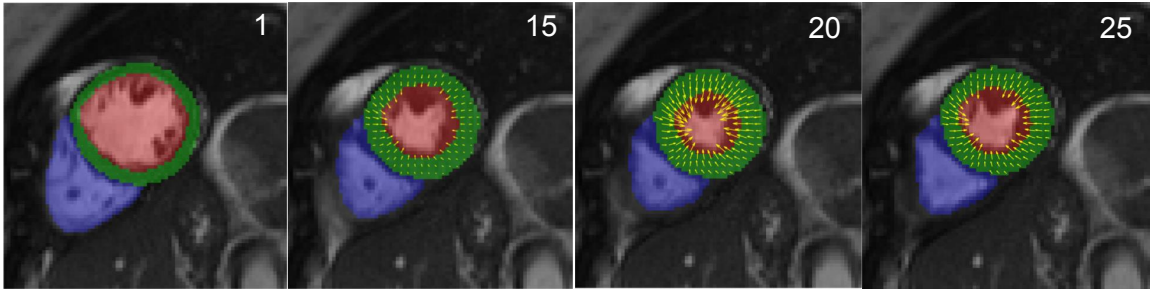


Figure 10.2: Visualisation results for simultaneous prediction of motion estimation and segmentation. Myocardial motions are from ED to other time points. Images are cropped to the region of interest for better visualisation.

Furthermore, the proposed joint method is capable of predicting a sequence of estimated motion fields and segmentations simultaneously. Here we show a visualisation result of the network predictions with segmentations and motions combined on frames in a cardiac cycle in Fig. 10.2. Myocardial motion indicated by the yellow arrows were established between ED and other time frames. Note that the network predicts dense motion fields, while for better visualisation, we only show a sparse representation around myocardium. To further validate the proposed unified model in terms of the motion estimation, Fig. 10.3(a)(b) shows a labeling results of the LV and RV boundaries along temporal dimension, which is obtained by warping the labeled segmentations available in ED frame to other time points, and Fig. 10.3(c) calculated the transformed LV volume over the cardiac cycle. These show that the proposed model is able to produce an accurate estimation, which is also smooth and consistent over time.

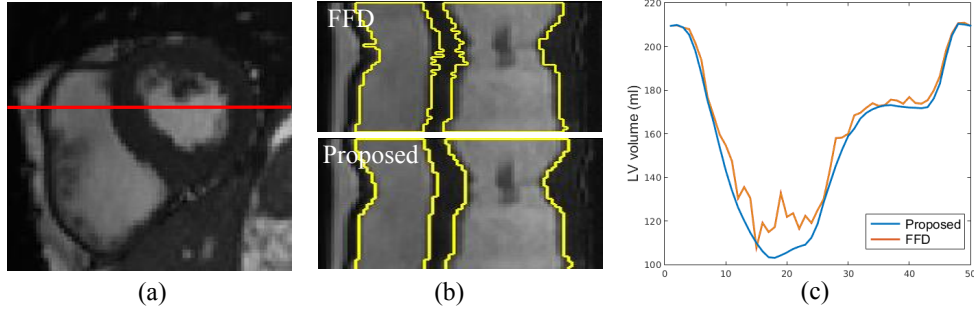


Figure 10.3: (a) (b) Labeling results obtained by warping the ED frame segmentation to other time points using FFD and the proposed joint model. Results are shown in temporal views of the red short-axis line. (c) Left ventricular volume (ml) of the subject by warping the ED frame segmentation to other time points in a cardiac cycle.

10.3 Joint Motion Estimation and Segmentation from Undersampled Cardiac MRI

10.3.1 Methods

Here we aim to predict the simultaneous motion estimation and segmentation directly from undersampled cardiac MR images and make sure that such predictions are as accurate and efficient as possible. An extension of the effective unified model (Motion-Seg Net) proposed in Section 10.2 is adapted to the application for undersampled MR data. Similarly, the proposed network architecture consists of two branches which perform motion estimation and segmentation jointly, and a well-trained sub-network for fully-sampled images is incorporated to provide additional supervision during the training process. Note that at test stage, only the undersampled sub-network is needed, and no fully-sampled data is required. The overall architecture of the model is shown in Fig. 10.4.

Unsupervised Cardiac Motion Estimation from Undersampled MR Image

Inspired by the success of the joint prediction network proposed in Section 10.2 which effectively learns useful representations, here we propose to adapt the network to undersampled MR data. In contrast to the fully-sampled case where only self-supervision is required for the motion

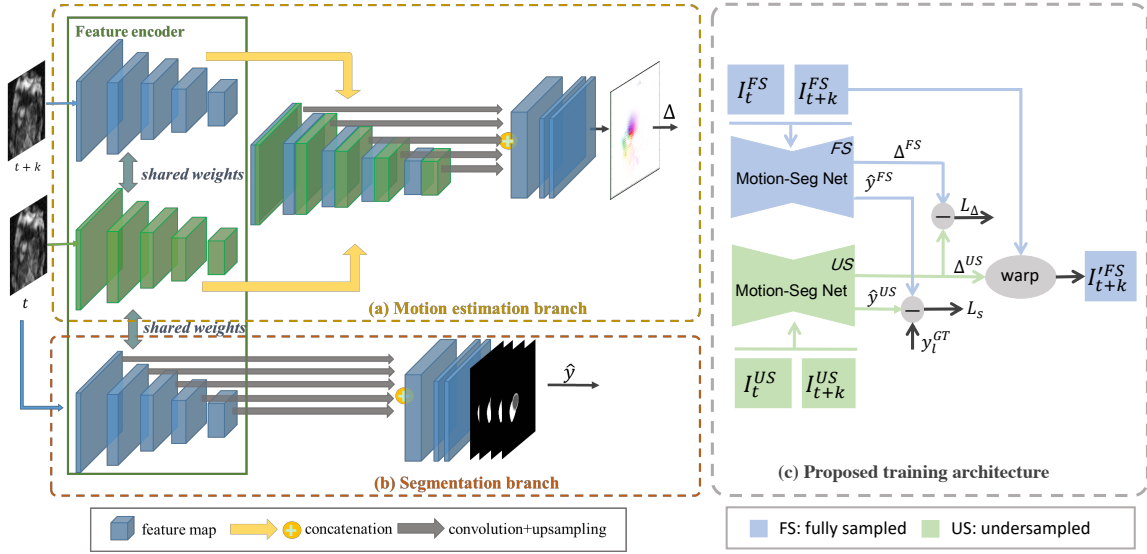


Figure 10.4: The overall schematic architecture of proposed network for joint estimation of cardiac motion and segmentation directly from undersampled data. (a) (b) The Motion-Seg net adopted from [194]. (c) Proposed architecture for training the Motion-Seg net on undersampled data. US: undersampled, FS: fully-sampled.

estimation, it is difficult for the undersampled images to merely rely on self-supervision, i.e., the intensity difference, due to the noises caused by aliased patterns. To address this, we propose to incorporate their corresponding fully-sampled image pairs as an additional supervision to guide the training for the undersampled images, and a schematic illustration of the model is shown in Fig. 10.4(a)(c).

The task is to find an optical flow representation between the target undersampled frame I_t^{US} and the source undersampled frame I_{t+k}^{US} , where the output is a pixel-wise 2D motion field Δ^{US} representing the displacement in x and y directions. We exploit a modified version of the network proposed in Section 10.2 for the representation learning, in which it mainly consists of three components: a Siamese network for the feature extraction of both target frame and source frame where the encoder is adapted from VGG-16 Net; a multi-scale concatenation of features from pairs of frames motivated by the traditional multi-level registration method [209]; and a bilinear interpolation sampler that warps the source frame to the target one by using the estimated displacement field $\Delta^{US} = (\Delta^{US}x, \Delta^{US}y; \theta_{\Delta}^{US})$, where the network is parameterised by θ_{Δ}^{US} which is learnt directly from undersampled MR data.

Due to the severe aliased patterns existing in the undersampled MR images, it is not practical to train the spatial transformer network purely based on minimising the intensity difference between the transformed undersampled frame and the target undersampled frame. To address this, we propose to introduce the fully-sampled image pairs as a supervision for the training. Specifically, instead of warping the undersampled source image, here we propose to transform the corresponding fully-sampled source image, which can be expressed as $I'_{t+k}{}^{FS}(x, y) = \Gamma\{I_{t+k}^{FS}(x + \Delta_{t+k}^{US}x, y + \Delta_{t+k}^{US}y)\}$. Then the network can be trained by optimising the pixel-wise mean squared error between I_t^{FS} and $I'_{t+k}{}^{FS}$. To ensure local smoothness, we maintain the regularisation term for the gradients of displacement fields which uses an approximation of Huber loss proposed in [43, 194], namely $\mathcal{H}(\delta_{x,y}\Delta^{US}) = \sqrt{\epsilon + \sum_{i=x,y}(\delta_i\Delta^{US}i^2 + \delta_y\Delta^{US}i^2)}$, where $\epsilon = 0.01$. Therefore, the loss function can be described as follows:

$$\mathcal{L}_m = \frac{1}{N_s} \sum_{(I_t, I_{t+k}) \in S} [\|I_t^{FS} - I'_{t+k}{}^{FS}\|^2 + \alpha \mathcal{H}(\delta_{x,y}\Delta_{t+k}^{US})], \quad (10.3)$$

where N_s stands for the number of sample pairs in the training set S , and α is a regularisation parameter to trade off between image dissimilarity and local smoothness.

However, it is observed that for heavily undersampled images, such weak supervision in Eq. 10.3 is not sufficient. Therefore, in order to push the learning results from undersampled data to be as accurate as that from fully-sampled data, we additionally introduce a pixel-wise mean squared error loss on the displacement fields between the estimation from undersampled data (Δ_{t+k}^{US}) and that from fully-sampled one (Δ_{t+k}^{FS}). Since only the motion of anatomical structures is of interest, here we propose to mask the region of interests (ROI) by utilising the predicted segmentation maps from fully-sampled data to allow that only errors from ROI can be backpropagated to contribute to the learning. The proposed loss term can be expressed as $\mathcal{L}_{\Delta_{t+k}} = \|(\Delta_{t+k}^{US} - \Delta_{t+k}^{FS}) * \mathbf{M}_t\|^2$, where $\mathbf{M}_t$ is a one-hot mask (1 for ROI, and 0 for background) generated from the segmentation maps from frame t of fully-sampled images. Thus, the overall loss function for motion estimation is as follows:

$$\mathcal{L}_m = \frac{1}{N_s} \sum [\|I_t^{FS} - I'_{t+k}{}^{FS}\|^2 + \alpha \mathcal{H}(\delta_{x,y}\Delta_{t+k}^{US}) + \beta \|(\Delta_{t+k}^{US} - \Delta_{t+k}^{FS}) * \mathbf{M}_t\|^2], \quad (10.4)$$

in which an additional trade-off parameter β is introduced. Note that no ground truth displacement fields are required during the training, thus the motion is still estimated unsupervisedly.

Joint Cardiac Motion Estimation and Segmentation from Undersampled MR Image

As introduced in Section 10.2, motion estimation and segmentation tasks are complementary. Similarly, here we couple both tasks for the joint prediction from undersampled MR data. The schematic architecture of the unified model is shown in Fig. 10.4.

The joint model consists of two branches: the motion estimation branch proposed in Section 10.2 which introduces additional supervision from fully sampled images, and the segmentation branch based on the network proposed in [21], where both branches share the joint feature encoder (Siamese style network) as shown in Fig. 10.4. As images are only temporally sparse annotated, predictions from corresponding fully-sampled images are used as supervision for those unlabelled data. Therefore a categorical cross-entropy loss $\mathcal{L}_s = -\sum_{l \in L} y_l^{GT} \log(f(x_l; \Theta^{US})) - \sum_{n \in U} \hat{y}_n^{FS} \log(f(x_n; \Theta^{US}))$ on labelled data set L and unlabelled data set U is used for segmentation branch, in which we define x_l and x_n as the input data, y_l^{GT} as the ground truth, $\hat{y}_n^{FS}$ is predictions from fully-sampled images and f is the segmentation function parameterised by Θ^{US} . Different from the loss function as stated in Section 10.2, here we do not employ the loss $\mathcal{L}_w$ between the warped segmentation and the target, as we find that for undersampled cases, minimising $\mathcal{L}_w$ could introduce more noises and uncertainties into the network training presumably because of the less accurate predictions. We empirically observed that this could lead to a small performance degradation especially for the segmentation branch.

As a result, the overall loss function for the joint model can be defined as:

$$\mathcal{L} = \mathcal{L}_m + \lambda \mathcal{L}_s, \quad (10.5)$$

where λ is a trade-off parameter for balancing these two tasks. $\mathcal{L}_m$ can be of the form of Eq. 10.3 or Eq. 10.4, and we will examine their comparisons in experiments.

10.3.2 Experiments and Results

We employed the same set of dataset as used in Section 10.2 for the experiments. Since only magnitude images are available, here we employed a phase map synthesis scheme proposed in [216] to synthetically generate phase maps (smoothly varying 2D sinusoid waves), in order to convert magnitude images to complex valued images and to make the simulation more realistic. In experiments, the synthesised complex valued images were back-transformed to regenerate k -space samples. The input undersampled images were generated by randomly undersampling the k -space samples using uniform radial undersampling patterns, as it generates images with better quality compared with Cartesian undersampling. For pre-processing, all training images were cropped to the same size of 192×192 , and intensity was normalised to the range of $[0,1]$. Parameters used in the loss function were set to be $\alpha = 0.001$, $\beta = 1$, and $\lambda = 0.01$, which were chosen via validation set. Fully-sampled sub-network parameters were loaded from [194], and we train the undersampled network using Adam optimiser with a learning rate of 0.0001. Data augmentation was performed on-the-fly, with random rotation, translation, and scaling.

As Section 10.2 has already shown that the joint model can significantly outperform model with single branch, here we mainly focus on the evaluation of the performance on undersampled data. We first evaluated the performance of motion estimation by comparing the proposed model with FFD algorithm [209], and the results are shown in Table 10.3. Here we examined the effect of different losses on the model's performance, where we termed method using $\mathcal{L}_m$ with the form of Eq. 10.3 as Proposed-A, and the one using Eq. 10.4 as Proposed-B. Motion fields were estimated between ES and ED frame, and mean contour distance (MCD) and Hausdorff distance (HD) were computed between the warped ES segmentations and ED segmentations. Results on fully-sampled (FS) images are presented in Table 10.3 as a reference. It can be observed that proposed methods consistently outperform FFD on all acceleration rates with $p \ll 0.001$ using Wilcoxon signed rank test, and is able to produce results that are close to the fully-sampled images. Furthermore, it can also be noticed that for higher acceleration rates ($6\times$ and $8\times$), Proposed-B produces significantly better results than Proposed-A ($p \ll 0.001$). This is reflected by the fact that higher undersampling rates result in more aliased images, therefore a

Table 10.3: Evaluation of motion estimation accuracy for undersampled MR data with different acceleration factors in terms of the mean contour distance (MCD) and Hausdorff distance (HD) in mm (mean and standard deviation). Loss function using $\mathcal{L}_m$ (Eq. 10.3) is termed as Proposed-A, and the one using $\mathcal{L}_m$ (Eq. 10.4) is termed as Proposed-B. Bold numbers indicate the best results for different undersampling rates.

Method		MCD			HD		
		LV	Myo	RV	LV	Myo	RV
FS	FFD	1.83 (0.53)	2.47 (0.74)	3.53 (1.25)	5.10 (1.28)	6.47 (1.69)	12.04 (4.85)
	Joint Model	1.30 (0.34)	1.19 (0.26)	3.03 (1.08)	3.52 (0.82)	3.43 (0.87)	11.38 (4.34)
3×	FFD	2.19 (0.49)	2.54 (0.74)	3.94 (1.38)	6.27 (1.64)	6.62 (1.72)	13.92 (5.03)
	Proposed-A	1.32 (0.40)	1.23 (0.31)	3.41 (1.22)	3.53 (0.89)	3.59 (1.10)	12.69 (4.47)
	Proposed-B	1.37 (0.45)	1.23 (0.31)	3.44 (1.22)	3.59 (0.98)	3.55 (1.10)	12.69 (4.45)
6×	FFD	2.80 (0.77)	2.74 (0.75)	4.48 (1.46)	7.83 (2.30)	7.26 (2.26)	15.63 (5.19)
	Proposed-A	2.10 (0.80)	1.44 (0.38)	3.84 (1.27)	4.79 (1.40)	3.98 (1.26)	13.45 (4.49)
	Proposed-B	1.74 (0.68)	1.34 (0.35)	3.68 (1.27)	4.20 (1.30)	3.77 (1.21)	13.08 (4.49)
8×	FFD	3.29 (0.97)	3.09 (0.99)	4.94 (1.67)	9.40 (2.70)	8.48 (3.05)	17.16 (5.75)
	Proposed-A	2.30 (0.97)	1.52 (0.46)	4.02 (1.37)	5.19 (1.71)	4.16 (1.32)	13.79 (4.60)
	Proposed-B	1.79 (0.70)	1.44 (0.39)	3.76 (1.30)	4.36 (1.40)	3.97 (1.28)	13.27 (4.55)

relatively strong supervision ($\mathcal{L}_\Delta$) is more needed to guide the learning in comparison to images with less aliasing (3×). In addition, the motion estimation performance can heavily rely on the chosen undersampling patterns. Model performance can be limited when using Cartesian undersampling pattern due to its strong aliasing artefacts under the same acceleration factor.

We further evaluated the segmentation performance of the model on undersampled data with different acceleration factors. Results reported in Table 10.4 are Dice scores computed with manual annotations on LV, Myo, and RV, as well as the clinical parameter ejection fraction (EF). It has been observed that Proposed-A and Proposed-B did not differ significantly in terms of segmentation performance, so here we only report results obtained from Proposed-B in Table 10.4. It can be seen that though there is a relatively small drop of performance as acceleration factors increase, the network is robust to train on undersampled data, and the clinical parameter predicted directly from undersampled data is very close to that from fully-sampled images. Furthermore, a visualisation result of the network predictions on 8× accelerated data in a cardiac cycle is shown in Fig. 10.5, where myocardial motion indicated by the yellow arrows were established between ED and other time frames. Overall, predictions directly from undersampled MR data are reasonably accurate, despite some small underestimations.

Table 10.4: Evaluation of segmentation performance under different acceleration factors in terms of Dice Metric (mean and standard deviation) and average percentage (%) error for ejection fraction (EF) compared with fully-sampled data.

Acceleration	LV	Myo	RV	EF
FS	0.9348 (0.0408)	0.8640 (0.0295)	0.8861 (0.0453)	-
3×	0.9303 (0.0450)	0.8596 (0.0309)	0.8884 (0.0433)	2.68%
6×	0.9214 (0.0475)	0.8424 (0.0310)	0.8804 (0.0456)	3.56%
8×	0.9141 (0.0487)	0.8260 (0.0343)	0.8658 (0.0523)	4.16%

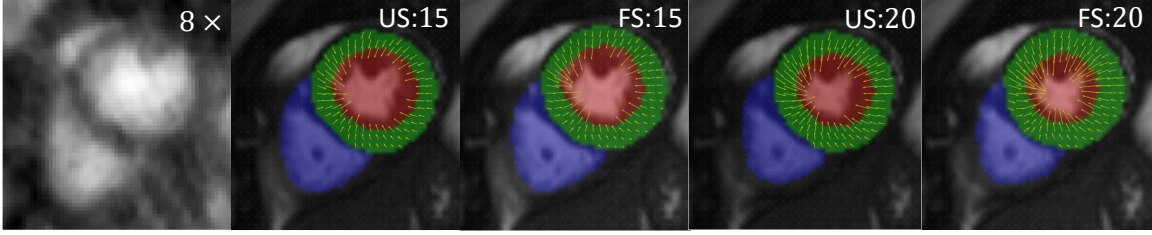


Figure 10.5: Comparison visualisation results for simultaneous prediction of motion estimation and segmentation on data with undersampling rates 8. Myocardial motions are from ED to other time points (numbers on the top right). Segmentations are overlaid on fully-sampled data. Images are cropped to the region of interest for better visualisation.

10.4 Conclusions

In this chapter, we presented a novel deep learning model for joint motion estimation and segmentation of cardiac MR image sequence, and its extension on joint prediction directly from undersampled MR image bypassing the usual MR image reconstruction stage. The proposed architecture is composed of two branches: a proposed unsupervised Siamese style spatial transformer network for motion estimation and a segmentation branch based on a fully convolutional network. A joint feature encoder is shared between the two branches, which enables the effective feature learning via multi-task training and also the weakly-supervised segmentation in terms of the temporally sparse annotated data. For predictions directly from undersampled images, a parallel well-trained sub-network for corresponding fully-sampled MR image pairs is introduced as a supervision source for training undersampled data, in order to push the predictions from undersampled data to be as accurate as possible. Both of these experimental results showed improvements of proposed models against baseline approaches in terms of accuracy and speed.

Chapter 11

Unsupervised Multi-modal Deformable Image Registration

This chapter is based on:

- **C. Qin**, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen. Unsupervised Deformable Registration for Multi-Modal Images via Disentangled Representations. IPMI (2019)

11.1 Introduction

Different medical image modalities, such as MRI, CT, and PET, show unique tissue features at different spatial resolutions. In clinical practice, multiple image modalities must often be fused for diagnostic or interventional purpose, providing the combination of complementary information. However, images from different modalities are often acquired with different scanners and at different time points with some intra-patient anatomical changes. It is of great importance to register multi-modal images for an accurate analysis and interpretation.

Multi-modal image registration is a challenging problem, due to the unknown and complex relationship between the intensity distributions of the images to be aligned. Also, there could be presence of features in one modality but missing in another. Previous multi-modal image

approaches either rely on information theoretic measures such as mutual information or on landmarks being identified in both images. However, information theoretic measures often ignores spatial information, and anatomical landmarks cannot always be localised in both images.

In this chapter, we propose a novel unsupervised registration method for aligning intra-subject multi-modal images, without the need of ground truth deformation fields, aligned multi-modal image pairs or any anatomical landmarks during training. To address this, our main idea is to learn a parameterised registration function via reducing the multi-modal registration problem to a mono-modal one in latent embedding space. In particular, our method decomposes images into a domain-invariant latent shape representation and a domain-specific appearance code based on the multi-modal unsupervised image-to-image translation framework (MUNIT) [117]. With the assumption that the intrinsic shape deformation between multi-modal image pairs is preserved in the domain-invariant shape space, we propose to learn an unsupervised diffeomorphic registration network directly based on the disentangled shape representations. A similarity criterion thus can be defined in the latent space, minimising the latent shape distance between warped moving image and target one. Additionally, a complimentary learning-based similarity metric is also proposed, which is defined via an adversarial loss to distinguish whether a pair of images are sufficiently aligned or not in the image domain. Since transformation is learned from a domain-invariant space, the method is directly applicable to bi-directional multi-modal registration without extra efforts.

Our main contributions can be summarised as follows: First, we present a learning-based unsupervised multi-modal deformable image registration method that does not require any aligned image pairs or anatomical landmarks. Second, we propose to learn a **bi-directional** registration function based on disentangled shape representation by optimising the proposed similarity criterion defined on both latent and image space. Third, we demonstrate that our proposed methods are competitive to state-of-the-art multi-modal image registration solutions in terms of accuracy, and have a much faster speed.

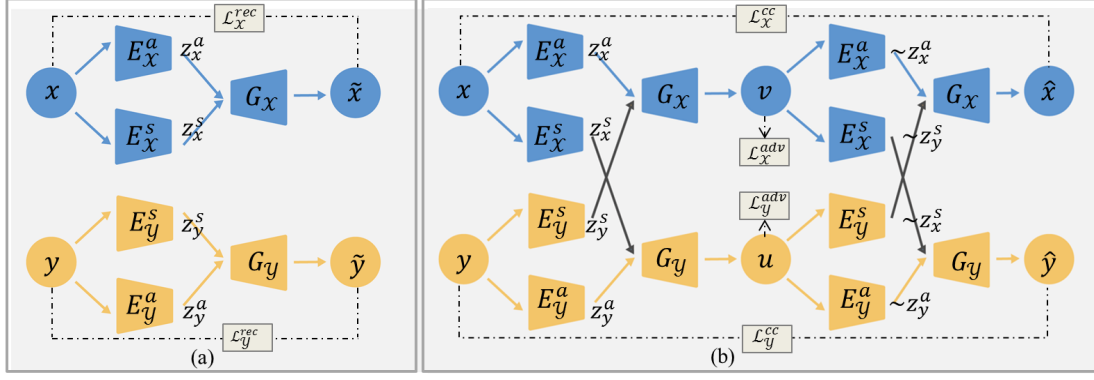


Figure 11.1: Image-to-image translation framework (Section 11.2.1). x and y are sample images in $\mathcal{X}$ and $\mathcal{Y}$ domain respectively. $\{E_x^s, E_y^s\}$ and $\{E_x^a, E_y^a\}$ are corresponding shape encoders and appearance encoders. $\{G_x, G_y\}$ are image generators. (a) Image self-reconstruction with $\{\tilde{x}, \tilde{y}\}$ reconstructed from $\{x, y\}$. (b) Image-to-image translation and cross-cycle reconstruction. $\{u, v\}$ are translated images from $\{x, y\}$ to domain $\{\mathcal{Y}, \mathcal{X}\}$ respectively. $\{\hat{x}, \hat{y}\}$ are cross-cycle reconstructed images. Black dotted lines and tilde notation indicate the consistency between variables.

11.2 Proposed Method

The goal is to learn a multi-modal deformable registration network in a fully unsupervised manner: without ground truth deformation fields, anatomical landmarks, or aligned multi-modal image pairs for training. We achieve this by embedding images of different modalities into a domain-invariant space via image disentangling, where any meaningful geometrical deformation can be directly derived in the latent space. Our method mainly consists of three parts: image disentangling network via unpaired image-to-image translation, a deformable registration network in the disentangled latent space and an adversarial network that implicitly learns a similarity metric in image space. A schematic illustration of our method is shown in Fig. 11.1 and Fig. 11.2.

11.2.1 Image disentangling via unpaired image-to-image translation

Huang et al. [117] and Lee et al. [153] have proposed to solve unpaired image-to-image translation problem through disentangled image representations, where images are embedded into a domain-invariant attribute space and a domain-specific attribute space, as shown in

Fig. 11.1. As described and shown in [117], domain-invariant attribute mainly captures the underlying spatial structure, and domain-specific attribute corresponds to the rendering of structure that is determined by imaging physics in our application. This approach formed the basis of our work. We briefly describe its main concept below that is related to our following registration work.

Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ denote unpaired images from two different domains, or in our application, two different imaging modalities. As illustrated in Fig. 11.1, image x is disentangled into a shape (content) code z_x^s in a domain-invariant space $\mathcal{S}$ and an appearance code z_x^a in a domain specific space $\mathcal{A}_\mathcal{X}$, where $E_\mathcal{X}^s$ and $E_\mathcal{X}^a$ encode x to z_x^s and z_x^a respectively. The generator $G_\mathcal{X}$ generates images conditioned on both shape and appearance vectors. Image-to-image translation is performed by swapping the latent codes in two domains, such as $v = G_\mathcal{X}(z_x^a, z_y^s)$ so that image y is translated to target domain $\mathcal{X}$. To train the framework for image-to-image translation and achieve representation disentanglement, a bidirectional reconstruction loss is used which comprises image self-reconstruction loss and latent reconstruction loss, i.e.,

$$\mathcal{L}_\mathcal{X}^{rec} = \mathbb{E}_x[||G_\mathcal{X}(E_\mathcal{X}^s(x), E_\mathcal{X}^a(x)) - x||_1], \quad (11.1a)$$

$$\mathcal{L}_{\mathcal{X}^s}^{lat} = \mathbb{E}_{x,y}[||E_\mathcal{Y}^s(G_\mathcal{Y}(z_x^s, z_y^a)) - z_x^s||_1], \quad (11.1b)$$

$$\mathcal{L}_{\mathcal{Y}^a}^{lat} = \mathbb{E}_{x,y}[||E_\mathcal{Y}^a(G_\mathcal{Y}(z_x^s, z_y^a)) - z_y^a||_1]. \quad (11.1c)$$

In order to better preserve the shape information, we also propose to incorporate an extra loss term to ensure cross-cycle consistency [153] :

$$\mathcal{L}^{cc} = \mathcal{L}_\mathcal{X}^{cc} + \mathcal{L}_\mathcal{Y}^{cc} = \mathbb{E}_{x,y}[||G_\mathcal{X}(E_\mathcal{Y}^s(u), E_\mathcal{X}^a(v)) - x||_1 + ||G_\mathcal{Y}(E_\mathcal{X}^s(v), E_\mathcal{Y}^a(u)) - y||_1]. \quad (11.2)$$

Besides, adversarial losses $L_\mathcal{X}^{adv}$ and $L_\mathcal{Y}^{adv}$ are employed to match the distribution of translated images to the image distribution in the target domain. Overall, our image-to-image translation network is trained by a weighted sum of image self-reconstruction loss, latent representation reconstruction loss, adversarial loss and the cross-cycle consistency loss. In addition, to avoid the trivial solution of having latent shape representations mapping to zero, here the appearance code is designed to be an eight dimensional vector, which enforces that the majority of content

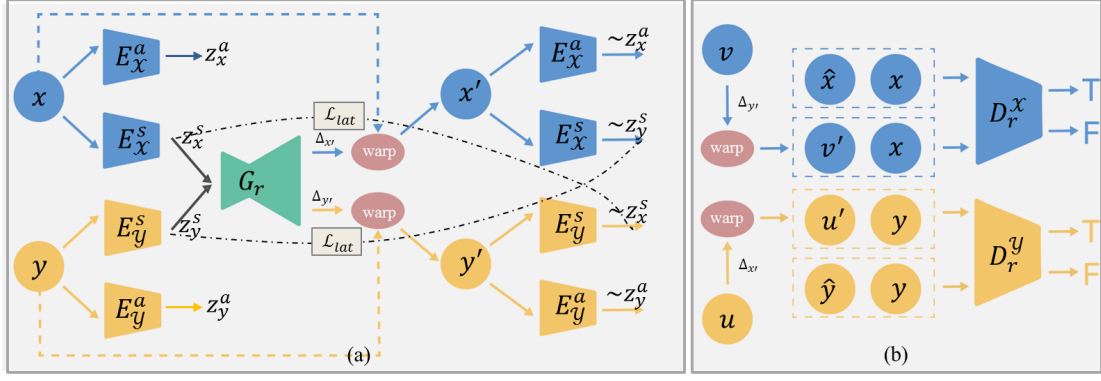


Figure 11.2: Overview architecture of the proposed models. (a) Multi-modal image registration via disentangled representations (Section 11.2.2). x' and y' are warped images from x and y . (b) Learning-based similarity metric in image space. u, v and $\hat{x}, \hat{y}$ are translated and reconstructed images respectively adopted from Fig.11.1. (a)+(b) Multi-modal image registration via combined metrics (Section 11.2.3). G_r is the registration network in latent space. D_r^x, D_r^y are discriminators in image space.

information should be contained in the shape representations. For more details, please refer to [117, 153].

11.2.2 Multi-modal image registration via disentangled representations

With image-to-image translation and disentangled attributes being learnt, we are able to reduce multi-modal registration problem to a mono-modal one by embedding images onto the domain-invariant latent space and learn the deformation there. A explanatory figure of our proposed network is shown in Fig. 11.2(a).

Specifically, images from different modalities are disentangled into a shared shape space $\mathcal{S}$ and different appearance spaces $\mathcal{A}_x$ and $\mathcal{A}_y$ respectively. The latent shape representations z_x^s and z_y^s contain high-level structure information of images which is capable of restoring the original image by combining with the appearance code. Relying on this, we propose to learn a deformable registration network by aligning images via these disentangled shape representations. When registering a moving image $y \in \mathcal{Y}$ to a fixed image $x \in \mathcal{X}$, the structure of the warped moving image $y' \in \mathcal{Y}$ should be close to that of the fixed one while keeping the

appearance unchanged. Therefore, a similarity criterion for training the registration network can be defined in the disentangled latent shape space. Specifically, we propose to learn a diffeomorphic registration network that receives latent shape representations as inputs and predicts a velocity field w . Deformation Δ between moving and fixed images is defined as an exponential map with respect to the velocities: $\Delta = \exp(w)$, which is implemented by an exponentiation layer as proposed in [142]. The detailed architecture of the registration network is shown in Fig. 11.3, which follows the idea proposed in Section 10.2. To train the network, the warped image y' is then encoded back to the latent shape space, and thus similarity between shape representations $E_{\mathcal{Y}}^s(y')$ and z_x^s can be enforced. In addition, since both images are mapped to a common feature space (modality-independent space), the registration network learnt in this space is directly applicable to be **bi-directional**, i.e., for both $y \rightarrow x$ and $x \rightarrow y$ registration. This is superior to learning a registration network in image space, which normally requires separate training for each direction. Therefore, by incorporating the bi-directional registration, the network can be trained by minimising the following similarity metric that is defined on latent space:

$$\mathcal{L}_{lat} = \mathbb{E}_{x,y} [||E_{\mathcal{Y}}^s(y') - z_x^s||_1 + ||E_{\mathcal{X}}^s(x') - z_y^s||_1] + \lambda_{\Delta} [\mathcal{H}(\nabla_{i,j}\Delta_{y'}) + \mathcal{H}(\nabla_{i,j}\Delta_{x'})], \quad (11.3)$$

where we penalise the gradients of the deformation fields $\Delta_{y'}$ and $\Delta_{x'}$ using an approximation of Huber loss [43] $\mathcal{H}(\nabla_{i,j}\Delta) = \sqrt{\epsilon + \sum_{m=i,j} (\nabla_i \Delta m^2 + \nabla_j \Delta m^2)}$ along both i and j directions to ensure the smoothness. λ_{Δ} is a regularisation parameter for a balance (trade-off) between different terms, and $\epsilon = 0.01$.

11.2.3 Multi-modal image registration via combined metrics

While disentangled latent shape representations can effectively capture high-level structural information, training with a latent similarity criterion only could possibly ignore some detailed structure deformations. To compensate this, we propose to combine the latent similarity criterion with an additional learning-based similarity metric in image space, as shown in Fig.11.2(b).

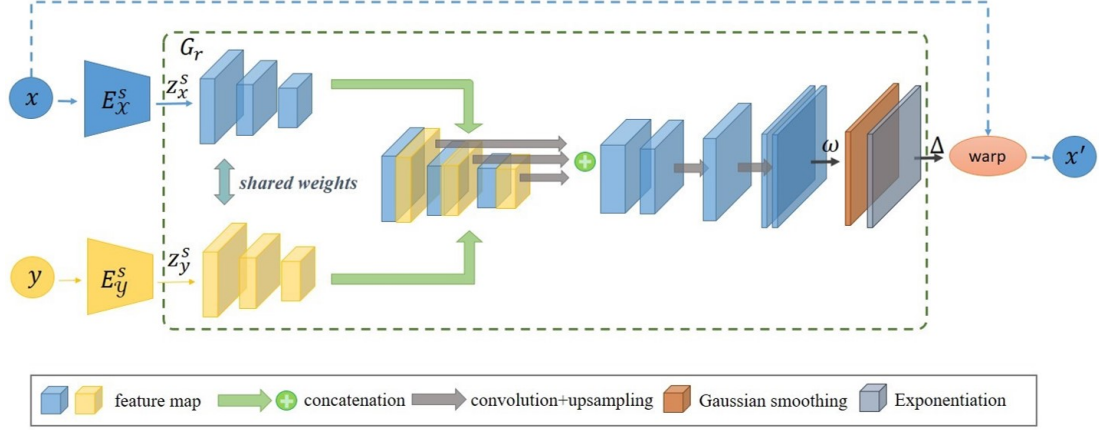


Figure 11.3: An illustrative architecture of the registration network in UMDIR method.

Similarly, here we define the learning-based similarity metric in image space also via image-to-image translation. However, during image-to-image translation, there could inevitably exist some mismatch of distributions between synthesized images and target images, especially when appearance distributions of real images are complex. Thus, mono-modal registration methods based on intensity similarities may not be sufficient. Therefore, instead of using a specific intensity-based similarity measure, similar to [80], we propose to learn a similarity metric function formulated by a patch GAN discriminator, which is trained to distinguish if a pair of image patches is well-aligned or not. Different from [80], to mitigate influence of distribution mismatch, we utilise the cross-cycle consistency of the translation network when designing the real pairs (well-aligned) and fake pairs (registered by network), i.e., $\{G_{\mathcal{X}}(E_{\mathcal{Y}}^s(u), E_{\mathcal{X}}^a(v)), x\}$ and $\{v', x\}$, where v' indicates the corresponding warped images of v . This is to enforce the discriminator to learn structure alignment instead of distribution differences. Architecture of discriminators follows the design of the feature encoder in registration network. Overall, we formulate the combined problem using the improved Wasserstein GAN (WGAN-GP)[99]: the image registration network G_r (generator) and two discriminators $D_r^{\mathcal{X}}$ and $D_r^{\mathcal{Y}}$ can be trained via alternatively optimising the respective composite loss functions:

$$\mathcal{L}_{D_r^{\mathcal{X}}} = \mathbb{E}_{\tilde{q} \sim \mathbb{P}_f} [D_r^{\mathcal{X}}(\tilde{q})] - \mathbb{E}_{q \sim \mathbb{P}_r} [D_r^{\mathcal{X}}(q)] + \lambda_{grad} \cdot \mathcal{L}_{grad}^{\mathcal{X}} \quad (11.4a)$$

$$\mathcal{L}_{D_r^{\mathcal{Y}}} = \mathbb{E}_{\tilde{p} \sim \mathbb{P}_f} [D_r^{\mathcal{Y}}(\tilde{p})] - \mathbb{E}_{p \sim \mathbb{P}_r} [D_r^{\mathcal{Y}}(p)] + \lambda_{grad} \cdot \mathcal{L}_{grad}^{\mathcal{Y}} \quad (11.4b)$$

$$\mathcal{L}_{G_r} = -\mathbb{E}_{\tilde{q} \sim \mathbb{P}_f} [D_r^{\mathcal{X}}(\tilde{q})] - \mathbb{E}_{\tilde{p} \sim \mathbb{P}_f} [D_r^{\mathcal{Y}}(\tilde{p})] + \alpha \mathcal{L}_{lat}, \quad (11.4c)$$

where $D_r^{\mathcal{X}}$ and $D_r^{\mathcal{Y}}$ are two discriminators for the bi-directional registration to distinguish real pairs and fake pairs in $\mathcal{X}$ and $\mathcal{Y}$ domain. $\{q, \tilde{q}\}$ and $\{p, \tilde{p}\}$ are {real, fake} pairs sampled from $\mathcal{X}$ and $\mathcal{Y}$ respectively. $\mathcal{L}_{grad}^{\mathcal{X}}$ is the gradient penalty for the discriminator $D_r^{\mathcal{X}}$ which can be expressed as the form of $\mathcal{L}_{grad}^{\mathcal{X}} = \mathbb{E}_{\hat{q} \sim \mathbb{P}_{\hat{q}}} [(\|\nabla_{\hat{q}} D_r^{\mathcal{X}}(\hat{q})\|_2 - 1)^2]$ with $\hat{q}$ sampled uniformly between q and $\tilde{q}$, and the same with $\mathcal{L}_{grad}^{\mathcal{Y}}$. α is a parameter to balance between the learning-based image space similarity metric and the latent space similarity measure.

11.3 Experiments

11.3.1 Datasets

We used two datasets for evaluation: one with clinical meaningful deformations in single modality (COPDGene) and one with real well-aligned multi-modality images (BraTS). In this case, we can have control over image-to-image translation quality and predicted deformation respectively.

COPDGene¹ The COPDGene study is a multicenter observational study to analyze genetic susceptibility for the development of chronic obstructive pulmonary disease (COPD)[201]. High-quality, volumetric lung CT scans were acquired to capture full inspiration cycle of each subject using a standardised imaging protocol. CT scans were reconstructed with slice thicknesses of 0.625, 0.75, or 0.9 mm depending on the CT scanner manufacturer, with corresponding slice intervals of 0.625, 0.5, and 0.45 mm, respectively. In our experiment, 1000 subjects are randomly retrieved for evaluation, with end inspiration and expiration volumes being used to derive the underlying breathing motion. Each pair of volumes was rigidly pre-aligned, cropped, and down-sampled into a 3D volume with size of $128 \times 128 \times 128$ and resolution of 2.5mm. We randomly split the 1000 subjects into 800/100/100 for train/test/validation, and on each subject we extract middle 10 slices. To simulate a multi-modality image registration problem, we synthesized a new modality using an intensity transformation $\cos(I \cdot \pi)$ as proposed in

¹The COPDGene study (NCT00608764) was funded by NHLBI U01 HL089897 and U01 HL089856 and also supported by the COPD Foundation through contributions made to an Industry Advisory Committee comprised of AstraZeneca, Boehringer-Ingelheim, GlaxoSmithKline, Novartis, and Sunovion.

[275] followed by Gaussian blurring and intensity normalization. Deformations were estimated between end inspiration and expiration frames of real CT and synthesized images.

BraTS’17 The Brain Tumour Segmentation (BraTS) 2017 dataset is obtained from the MICCAI BraTS 2017 challenge [22, 177]. Specifically, it provides a large dataset of multi-modal MRI scans (native T1, T2, T2-FLAIR, and T1Gd) for patients with glioblastomas. Overall, the available training set consists of 285 cases, and for each case four image modalities were standardized into a 3D volume in size of $240 \times 240 \times 155$ with 1 mm isotropic resolution. In our experiments, we utilise the T1 and T2-weighted images to define a multi-modality dataset to demonstrate the effect of our proposed approach. The set is randomly split into 225/30/30 for train/validation/test, and central 20 slices of each subject were extracted. As provided T1 and T2 images are already aligned, we generated synthetic deformation fields by spatially transforming one of the modality (T1) using elastic transformations on control points followed by Gaussian smoothing. The synthetic deformation is only used as ground truth (GT) for evaluations.

11.3.2 Experimental Settings

Implementation Details: For image-to-image translation, we built our network based on MUNIT implementation with changes as discussed in Section 11.2.1. The network is trained using the default settings as in [117]. Our registration network adopts the same architecture as in Section 10.2 with an additional exponentiation layer [142] as the last layer, as shown in Fig. 11.3. In our implementation, we pre-train the image-to-image translation network using unpaired images, and then multi-modal registration and discriminator networks are trained. Our networks are implemented on PyTorch, using Adam optimiser for training with a learning rate of 0.0001. Hyper-parameters were chosen based on the performance on validation set, with $\lambda_{grad} = 10$ and $\lambda_{\Delta} = 1$. Run time reported for each method was tested on the same PC with 32G RAM, 3.6GHz CPU, and Quadro P4000 GPU.

Evaluation Measures: For COPDGene dataset, we evaluate the registration accuracy indirectly via provided lung segmentation masks, as no ground truth deformation fields are

available. Dice score, mean contour distance (MCD) and Hausdoff distance (HD) are computed between lung masks of fixed and warped moving images. For BraTS dataset, synthetic deformation fields are used as GT, thus pixel-wise root mean square error ($\text{RMSE}(\Delta)$) is calculated for the evaluation. Also, as pairs of aligned images are available, pixel-wise intensity error ($\text{RMSE}(I)$) can be calculated when transforming back the deformed image. Additionally, for both datasets, analysis of Jacobian matrix $J_{\Delta}(m) = \nabla(\Delta(m))$ were conducted on the dense deformation fields Δ of each pixel m . Gradients of $J_{\Delta}(m)$ (Grad Det-Jac) are calculated as a metric to show the smoothness of Jacobian. Besides, average run time of each method is reported.

Competing Methods: We compare with two well-established multi-modal image registration methods: A mutual information based approach using the **Elastix** toolbox [173] and the **MIND** approach [108]. Also, we compare with the diffeomorphic Demons method [254] which can deal with multi-modal images using the proposed learned image-to-image translation network. Specifically, we first translate the appearance of the moving image to that of the fixed image and then run the diffeomorphic Demons algorithm on the image pair. We term this method as **I2I+DiffDem**. A hierarchical multiresolution optimisation scheme was used for all. Parameters were determined via searching on the validation set of each dataset separately while considering both the registration accuracy and speed. Results reported are best performance we can achieve. In addition, as no other deep learning based unsupervised multi-modal registration has been proposed yet, here we employ a diffeomorphic extension of existing mono-modal registration method [80] that is enabled for multi-modal images similarly as I2I+DiffDem, termed as **I2I+GAN**. Its registration network architecture follows Section 11.2.2 where instead of estimating deformation Δ , velocity field is estimated to ensure the diffeomorphism. Its discriminator has also been adapted for multi-modality as discussed in Section 11.2.3. Finally, we term variants of our proposed method as **UMDIR-Lat**, **UMDIR-GAN** and **UMDIR-LaGAN**, corresponding to models shown in Fig.11.2 (a), (b) and (a)+(b).

Table 11.1: Evaluation of multi-modal registration on COPDGene dataset in terms of average Dice score, MCD, HD (unit: pixel), run time (GPU/CPU) and Grad Det-Jac ($\times 10^{-2}$) of each deformation field.

Method	Avg. Dice	MCD	HD	Grad Det-Jac	Time(s)
MIND [108]	0.9365 (0.0822)	1.168 (0.798)	11.129 (5.736)	4.52 (0.40)	-/12.09
Elastix [173]	0.9497 (0.0822)	0.934 (0.760)	9.970 (5.772)	2.29 (0.61)	-/105.66
I2I+DiffDem	0.9347 (0.0676)	1.274 (0.923)	11.477 (5.434)	6.14 (0.39)	2.45/3.96
I2I+GAN	0.9553 (0.0444)	0.912 (0.628)	9.383 (4.977)	2.28 (0.45)	0.12/2.81
UMDIR-GAN	0.9613 (0.0357)	0.819 (0.546)	9.188 (4.981)	2.19 (0.48)	0.07/1.62
UMDIR-Lat	0.9603 (0.0349)	0.823 (0.462)	8.469 (4.491)	2.73 (0.80)	0.07/1.62
UMDIR-LaGAN	0.9672 (0.0280)	0.710 (0.436)	8.257 (4.432)	2.79 (0.60)	0.07/1.62

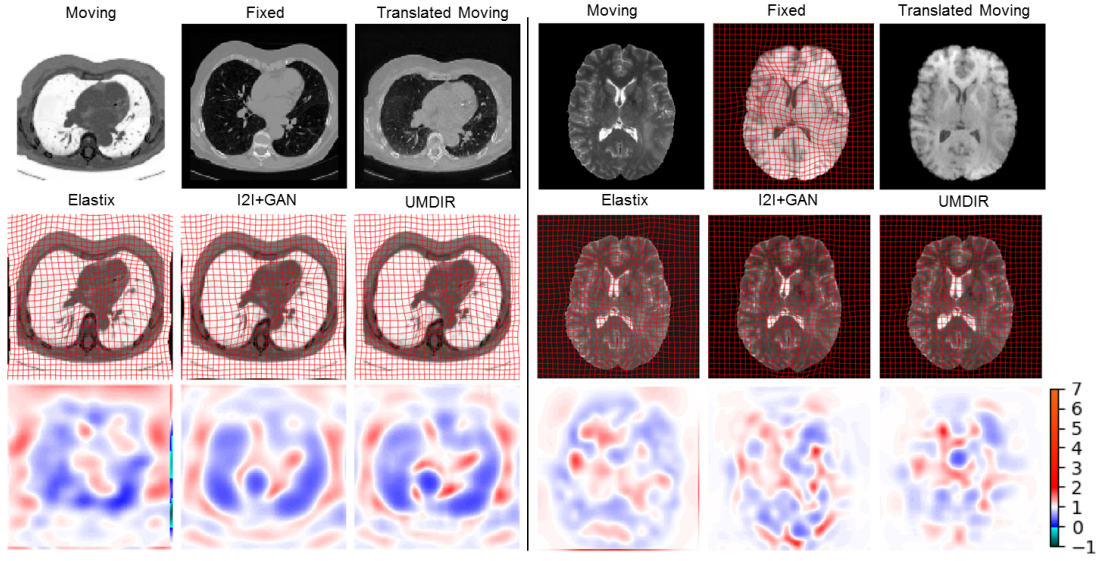


Figure 11.4: Visualisation results of our model against baseline methods, where warped moving images, corresponding estimated deformation fields and Jacobian determinant are shown. Left: COPDGene data; Right: BraTS data with GT overlaid on Fixed image.

11.3.3 Results

First, we evaluated the performance of our methods on COPDGene data. The quantitative results for this are shown in Table 11.1. The registration performance is evaluated via measuring the Dice score, MCD and HD between warped lung segmentation at expiration and the GT lung segmentation at inspiration, as well as the gradient of Jacobian determinant. It can be seen that compared to the baseline methods, both traditional multi-modal registration methods and image-to-image translation plus mono-modal registration methods, our proposed UMDIR methods outperforms them in terms of Dice, MCD and HD with Grad Det-Jac smaller than or

Table 11.2: Evaluation of bi-directional multi-modal registration on BraTS dataset in terms of RMSE(Δ) (unit: pixel) and Grad Det-Jac ($\times 10^{-2}$) for $T2 \rightarrow T1$ registration, and RMSE(I) for the inverse direction. Average run time (GPU/CPU) is also provided.

Method	$T2 \rightarrow T1$		$T1 \rightarrow T2$		Time(s)
	RMSE(Δ)	Grad Det-Jac	RMSE(I)	Grad Det-Jac	
MIND [108]	1.266 (0.253)	3.58 (0.25)	0.045 (0.011)	3.60 (0.24)	-/18.98
Elastix [173]	1.260 (0.225)	1.23 (0.14)	0.089 (0.013)	1.22 (0.17)	-/66.97
I2I+DiffDem	1.391 (0.183)	0.83 (0.13)	0.057 (0.015)	0.87 (0.14)	0.81/4.44
I2I+GAN	1.250 (0.218)	1.30 (0.11)	0.074 (0.014)	2.10 (0.43)	0.20/6.44
UMDIR-GAN	1.202 (0.196)	1.35 (0.14)	0.067 (0.010)	1.88 (0.32)	0.08/3.60
UMDIR-Lat	1.146 (0.232)	1.04 (0.14)	0.067 (0.012)	1.37 (0.41)	0.08/3.60
UMDIR-LaGAN	1.126 (0.214)	0.97 (0.11)	0.064 (0.010)	1.05 (0.20)	0.08/3.60

in the same range with other competing methods. In particular, success of training with latent similarity criterion implies that the learned domain-invariant attribute is capable of extracting and preserving intrinsic shape feature that is informative enough to guide the geometrical transformation for registration. In addition, Fig. 11.4 displays the warped moving images along with their corresponding deformation fields and Jacobian determinant, where it can be observed that our proposed model is able to achieve high accuracy with smooth and regular deformation fields. Furthermore, by combining both latent similarity and cross-cycle adversarial similarity metrics, we see a further improvement of performance on the COPDGene dataset in terms of accuracy, which indicates that these two metrics could be complementary. In terms of registration speed, our proposed methods are significantly faster than traditional baseline methods. In particular, the UMDIR methods are faster than I2I+GAN, as they learn the registration network directly from latent representations, bypassing the image-to-image translation stage. Note that the translated moving images in Fig. 11.4 are only used in the I2I+DiffDem and I2I+GAN.

Additionally, we also evaluated the bi-directional multi-modal registration performance on BraTS dataset, where ground truth deformation fields are available from $T2$ to $T1$, and ground truth aligned images are known for the $T1$ to $T2$ registration. Note that our proposed latent method can realise such bi-directional registration directly without any further training, while other competing methods need to be optimised or trained separately for both $T1 \rightarrow T2$ and $T2 \rightarrow T1$ directions. Quantitative results are shown in Table 11.2, with RMSE(Δ) and Grad

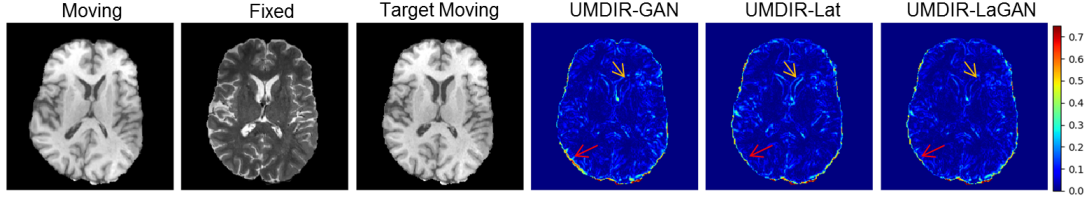


Figure 11.5: Comparison between variants of proposed methods for T1 to T2 registration. Error maps are pixel-wise intensity differences between warped images and target moving images. Red arrows indicate regions where UMDIR-Lat outperforms UMDIR-GAN, and yellow arrows point out regions where UMDIR-GAN outperforms UMDIR-Lat. Combined metrics shows better performance than each of them separately.

Det-Jac calculated for $T2 \rightarrow T1$, and RMSE(I) for $T1 \rightarrow T2$. An example of visualisation results is shown in Fig. 11.4. From both quantitative and qualitative results, it can be seen that the proposed UMDIR methods can achieve competitive registration performance compared with other competing methods and also with smooth and regular deformations. Though MIND achieved a lower RMSE(I) for $T1 \rightarrow T2$ registration, this is at the cost of separate training for each direction. Our proposed methods deliver a one-shot bi-directional solution in a noticeably higher processing speed. In addition, learning registration directly from disentangled latent representations bypasses the image-to-image translation stage, which can potentially avoid problems caused by the image translation quality such as image noises or hallucinations that can lead to inaccurate registration. This possibly explains the better performance of our proposed methods against I2I with mono-modal registration methods. To examine the differences of latent similarity criterion and cross-cycle adversarial similarity metric as well as the benefits of combined metrics, in Fig. 11.5, we compared the registration performance with different metrics in the inverse direction ($T1 \rightarrow T2$) by visualising the pixel-wise intensity RMSE. It can be observed that by combining these two complementary metrics, UMDIR-LaGAN is able to preserve the advantages of each metric and produces more accurate registration.

11.4 Conclusions

In this chapter, we have presented a novel deep learning based model for fully unsupervised multi-modal deformable image registration. The proposed models reduce the multi-modal registration problem to a mono-modal one via exploiting the disentangled latent embedding that

is learnt from an unpaired image-to-image translation framework. For training the registration network, we proposed a distance loss in latent shape space and a cross-cycle adversarial loss defined in image space as similarity metrics. Experimental results showed improvements of our proposed models against other conventional approaches in terms of both accuracy and speed.

Chapter 12

Conclusions

This thesis has focused on addressing several key challenges in the main components of medical imaging pipeline, including image reconstruction and analysis. A summary of thesis achievements and some potential future works are presented in the following.

12.1 Summary of Thesis Achievements

White Matter Hyperintensity Segmentation

The accurate assessment of WMH burden is of crucial importance for epidemiological studies to determine associations between WMHs, cognitive and clinical data, their causes, and the effects of new treatments in randomised trials. Manual delineation of WMHs is often time-consuming or impossible for large scale data set, and can often introduce inter- or intra-observer variability. To address this, Chapter 4 proposed an automatic WMH segmentation method based on supervised and semi-supervised large margin algorithms. The proposed model detects WMHs first via a global large margin classifier learnt across the training data, followed by fine-tuning using the proposed semi-supervised method on the target-specific data. This enables both the common information shared across different subjects and individual-specific information from the target subject to be utilised. The automatic segmentation results were compared with

WMH masks derived from expert annotations in terms of overlap, volumetric agreement and difference. Both the qualitative visualisation results and the quantitative scores showed the encouraging performance of the proposed model.

However, in many cases, stroke lesions often coexist with WMHs on patients with vascular diseases, and they also appear hyperintense on T2-weighted FLAIR MR images. In the assessment of WMH burden, it is important to exclude stroke lesions as they originate from different pathologies. Therefore, Chapter 5 further proposed an approach for the segmentation and differentiation of WMHs and stroke lesions. In particular, the proposed model is based on a CNN framework, termed uResNet, which has a similar U-net architecture but with residual block at each layer. Class imbalance between different types of lesions and normal tissues is considered in the framework, which has been examined via different loss functions and sampling strategies. Comparison results indicate that the proposed uResNet architecture outperforms other well established and state-of-the-art algorithms.

Dynamic Cardiac MR Image Reconstruction

Accelerating the data acquisition of dynamic MRI is a challenging ill-posed inverse problem, which has raised great interest from both the signal processing and machine learning community over the last decades. As is known, dynamic MRI sequences exhibit strong spatio-temporal redundancies between frames. To exploit such temporal redundancies, Chapter 7 presented an end-to-end deep learning solutions for the accelerated dynamic MR image reconstruction, based on a convolutional recurrent neural network, termed CRNN-MRI. The CRNN-MRI consists of a cascade of CRNN blocks implicitly learning iterative denoising interleaved by data consistency layers to enforce data fidelity. It models the recurrent correlations along two dimensions: both the temporal dimension and the iterative optimisation step, and two variants of convolutional recurrent units have been proposed. Experiments indicate the effectiveness of the proposed architecture against other competing methods.

In addition to exploit temporal redundancies in image domain, a novel approach named k - t NEXT was proposed in Chapter 8 which learns to iteratively recover the images by alternating

between the complementary x - f and image domains. In particular, k - t NEXT is composed of a xf -CNN that performs signal recovery in x - f domain and a CRNN-MRI network that reconstructs images in image domain, where these two networks were trained jointly. By exploiting k - t correlations in both x - f and image domains, it has shown improvements over the state-of-the-art dynamic MR reconstruction methods both quantitatively and qualitatively.

Image Registration and Motion Estimation

Medical image registration is a key component in medical image analysis. One of its applications is motion estimation in cardiac MRI, which is an essential step for the dynamic exploration of the cardiac function. In Chapter 10, a joint deep learning framework for the cardiac MRI segmentation and motion estimation was proposed, where it consists of two branches: an unsupervised spatial transformer network based branch for motion estimation between frames, and a segmentation branch based on a fully convolutional neural network. By sharing a joint feature encoder between these two branches, it enables effective feature learning via multi-task training, and experiments showed improvements of the proposed model against baseline approaches. Besides, extensions of the proposed framework on undersampled MRI data has also been investigated. In particular, a parallel well-trained sub-network for corresponding fully-sampled MR image pairs is introduced as a supervision source for training undersampled data, in order to ensure the predictions from undersampled data to be as accurate as possible. It has been shown that the network is robust to undersampled data, and results predicted directly from undersampled images are close to that from fully-sampled ones, which could potentially enable fast analysis for MR imaging.

On the other hand, multi-modal image registration is a challenging problem and has not yet been fully explored in deep learning. Chapter 11 has presented a novel deep learning method for unsupervised multi-modal deformable image registration, termed UMDIR. The proposed models reduce the multi-modal registration problem to a mono-modal one via exploiting the disentangled latent embedding that is learnt from an unpaired image-to-image translation framework. The multi-modal registration network is thereby learnt based on the latent shape representation

which is corresponding to the underlying structures of the data, and is trained by minimising a distance loss in latent shape space and a cross-cycle adversarial loss defined in image space. Experimental results showed competitive performance of the proposed models against other traditional approaches in terms of both accuracy and speed.

12.2 Limitations and Future Work

In this section, we discuss some limitations of the work presented in this thesis as well as some problems for machine learning in medical imaging in general. Then some discussions and suggestions for the future research directions are also provided. Note that the limitations and future directions are not limited to the discussions below.

12.2.1 2D Networks vs 3D Networks

The choice of using 2D or 3D networks has always been a topic of discussions. It is well known that 3D networks generally have more parameters and complexities than 2D networks, and thus 3D networks are more difficult and computationally expensive to optimise. Commonly the choice of 2D or 3D networks depends on the acquisition types of images as well as its applications. As discussed in Chapter 3, for instance, FLAIR images are often acquired as 2D images with large slice thickness and then stacked into a 3D volume, and manual delineation is also performed slice by slice. In this case, 2D networks are more natural choices considering the acquisition and annotation process in comparison to 3D networks, which can also largely reduce the computational cost. In addition, motion artefacts between different slices may also negatively affect 3D networks. For instance, in the work of [27], it was shown that a 3D U-net performed worse than a 2D U-net when evaluated on the automated cardiac diagnosis challenge (ACDC) dataset. Furthermore it was found that processing images in a slice-by-slice fashion was beneficial due to the relatively large slice thickness. In addition, for 2D sequence data, it is more efficient to employ 2D networks compared to 3D convolutions as shown in Chapter 7, where 2D convolutions at spatial and temporal dimensions separately can learn meaningful

variant features in contrast to invariant feature learning in 3D convolutions. It makes more sense here to use 2D convolutions as the resolution at spatial and temporal dimensions are very different.

Undoubtedly, it is also very important to consider the 3D volume context for image segmentation tasks. For example, when experts annotate organs, they may also need to consider the information at neighboring slices. Therefore, in order to take into account such 3D information as well as to preserve the benefits of 2D networks, several approaches using “2D+” networks have been proposed. Specifically, these methods include multi-view networks to utilise multi-planar information (i.e. coronal, sagittal, axial views) [180], multi-slice networks concatenating multiple slices along channel axis, and 2D networks combined with RNNs to incorporate context across multiple slices [188]. These networks inherit the advantages of 2D networks while still being capable of taking advantage of through-plane spatial context for segmentation with 3D consistency.

However, in applications such as volume registration, it is necessary to utilise 3D volume information to estimate 3D deformation fields between volumes. Therefore, in such circumstances, 3D networks may be a better choice if the data is also acquired in 3D. Chapter 11 has presented an unsupervised multi-modal image registration technique based on deep learning models, but it was demonstrated on 2D images. However, in practice, 3D image registration is of more clinical use, and thus extensions towards 3D volume registration are necessary. Future work on this aspect includes investigating efficient and effective networks for 3D applications, such as based on patch-based methods or the use of convolutions with reduced computations but large receptive field sizes.

12.2.2 Model Generalisation

One common limitation of learning-based algorithms is their lack of generalisation capabilities when presented with out-of-distribution samples. For instance, it is unclear how deep learning models trained on MR scans with specific parameter settings or anatomical structures can generalise well to other MR scans with different anatomy or scan parameters. In this

respect, non-learning based methods that mostly rely on iterative optimisation do not have such generalisation issue when performing in unseen domains compared to learning-based algorithms. However, there are also some challenges and limitations for non-learning based methods regarding this. As discussed in Chapter 6.1 for the MR image reconstruction, conventional non-learning based models such as compressed sensing methods often impose strong assumptions on the underlying data, which requires manual adjustments of hyperparameters depending on the application. Though non-learning based models may not be constrained by the data distributions, regularisation functions as well as their hyperparameters need careful selection when applying them on different data domains and this process is often non-trivial and task-specific. For instance, too strong sparsity or ℓ_1 penalties can lead to cartoon-like artefacts. In contrast, deep learning methods are able to exploit the prior knowledge that are learnt from the large amount of data, removing the need for manual adjustment. Similarly, for image registration, though optimisation-based methods such as FFD are applicable for various data distributions, they also require the adjustment of hyperparameters and the selection of regularisations for each specific task and data domain. The speed of these methods is often relatively slow due to requirement to use iterative algorithms for optimisation.

The limited out-of-distribution generalisation power of deep learning based models will inevitably prevent them to be deployed in real world applications. In contrast to conventional methods that rely on iterative optimisation, it is not practical to train a large number of separate deep learning models with respect to different parameter settings, which will also limit their translational potential and clinical use [140]. Therefore, an investigation of the model generalisation ability is essential. Domain adaptation can be one of the future research directions to improve the model performance on unseen domains with a few or even no additional labels. It aims to learn models from a source data distribution while ensuring that to also perform well on a different target data distribution. One possible direction is to synthesise data from unseen domain to the source domain in the appearance level or vice versa, so that models trained on one domain can be well adapted to the other. An alternative direction is to match distributions of different domains in feature space or adapt the model in a domain-invariant disentangled representation space. The work in Chapter 11 was based on the idea

of representation disentanglement, which can be readily extended for unsupervised domain adaptation. It would also be interesting to explore the combination of deep learning approaches with non-learning based methods, taking advantage of the effective deep representation learning to extract meaningful features while preserving the generalisation capability of non-learning based methods.

12.2.3 Learning with Limited Supervision

Most current state-of-the-art approaches are supervised learning methods which heavily depend on the training data and training labels. However, training labels in medical images are very expensive and time-consuming to obtain. In some cases, training labels are even impossible to acquire, such as the dense deformation fields for registering images, and in some other problems, it is not possible for humans to perform, such as image reconstruction and image super-resolution. However, there are abundant unannotated data available that are not utilised. Developing methods towards exploiting these large amount of unlabelled information is essential.

Weak supervision: Manual annotation of lesions is laborious and time-consuming, and even impossible for large scale dataset. Chapter 4 introduced a semi-supervised scheme to utilise both the limited annotated data and those large amount of unlabelled information. However, the performance of the method is inevitably inferior to fully supervised deep learning approaches. Alternatively, weak supervisions, such as bounding boxes, dot annotations or brush strokes, are relatively easy to acquire and can also be used as guidance for segmentation. As indicated in [28], point annotation takes 10 times less than a full annotation. Though there are some efforts in this direction, their performances are not yet comparable to fully supervised methods. It is interesting to research on unsupervised or weakly supervised methods for lesion segmentation, which would be clinically valuable and meaningful.

Self-supervised learning: One of the ways to learn useful representations without any human annotation is to explore the intrinsic data structure within data itself. Examples of self-supervisory signals include self-reconstruction, spatial context, temporal similarities or ordering, etc. Chapter 10 has presented to exploit the temporal changes as the supervisory signal for motion tracking, and Bai et al. [20] also exploited anatomical positions to learn features in a self-supervised manner. The success of these applications indicates the effectiveness of self-supervised representation learning, and it will be a promising direction to fully utilise those unlabelled information. Particularly, medical data is more structured, and such prior knowledge could be potentially exploited as a type of supervisory signal. In addition, information across different modalities could also be used, such as relating imaging data with non-imaging data to explore their intrinsic relations as supervisory signals.

Unsupervised representation disentanglement: As introduced in Chapter 11, representation disentanglement is an unsupervised learning technique that breaks down or disentangles features into each factor that could be encoded into separate dimensions of a vector representation, in order to better understand the explanatory factors behind observations. In the context of medical imaging, data can be disentangled into spatial anatomical factors and non-spatial modality or appearance factors, as shown in Chapter 11. Though it shows promising results, the degree of explicitness of the disentanglement still remain unclear. One interesting research direction is to incorporate some shape models or prior knowledge into the networks as constraints to ensure the disentanglement explicitness. Alternatively, it would also be interesting to investigate its joint training with some other relevant tasks for better representation learning. For instance, as presented in Chapter 11, it could be helpful to refine the image-to-image translation network during the training of the registration network, so that the domain-invariant attribute can be better enforced to be equivalent to the shape that could be represented via geometrical transformations.

12.2.4 Accelerated Dynamic Parallel Imaging

Research presented in Chapter 7 and Chapter 8 only considers a single coil setup. However, in practice, parallel imaging is now the default option to reduce MRI scan time, and nearly all clinical MRI scanners are equipped with parallel imaging technology [140]. Thus, it is necessary to investigate such methods in a scenario where multiple coil data from parallel MR imaging can be used jointly for higher acceleration acquisition. Currently, some deep learning techniques have been proposed for improving parallel imaging [140], but most of them focused on static image reconstruction. Exploiting temporal redundancies as well as multi-coil data information for even higher accelerated acquisitions is an interesting research direction to look into. Potential ways to extend from a single-coil setup to a parallel setup include reconstructing each coil image separately or reconstructing the sensitivity combined image with data consistency at each coil. It would also be interesting to explore the spatial correlations between multi-coil data together with exploiting the spatial-temporal redundancies of the temporal sequences in a multi-dimensional way. In addition, estimating sensitivity encoding matrix in a deep learning framework via attention mechanisms is also interesting. This could potentially further improve the reconstruction performance and also enable the end-to-end optimisation pipeline.

12.2.5 Class Imbalance

Large class imbalance in medical image segmentation is generally an issue that must be considered. Loss functions that take into account the class imbalance have the drawback that they have additional class weighting parameters to tune, and evaluation results actually indicated their performances similar to classical cross-entropy in practice. To address this issue, as shown in Chapter 5, a balanced patch sampling strategy was proposed to reduce the very large class imbalance in the data. Though the use of the proposed sampling strategy shows its effectiveness, a performance gap between WMH segmentation and stroke lesion segmentation still exists, where the dice score of stroke segmentation is much lower than that of WMH segmentation. This is partially due to the class imbalance between these two classes, where a lot of computational effort might have been spent on optimising to perform well in large and rela-

tively easy-to-segment sections of an image. Therefore, methods to address the class imbalance problem for better segmentation are still needed. Potential approaches could be to design some novel losses for unbalanced data or to rely on some generative models to generate synthesised data for balanced training.

12.2.6 Model Interpretability

Another concern about deep learning models is their lack of interpretability. Despite their good performance, interpreting deep learning models is extremely difficult. However, interpretation and explanation of models and their results are very important, especially in the field of medical imaging, where explanations associated with decision makings are needed. On the other hand, the lack of interpretability may also make the model unpredictable and unreliable. Recent studies have shown that adding a small perturbation to the images can result in the models' prediction failures. Thus building resilient algorithms robust to potential attacks remains an unsolved and challenging problem. One potential solution is to increase the model interpretability by mimicking the way that radiologists would interpret images. This could be achieved by estimating confidence maps such as attention maps to provide evidence for decisions or uncertainty maps to see where models are unsure. Alternatively, prior knowledge could be possibly incorporated into networks to guide the model interpretation and to inform human decision-making.

Bibliography

- [1] Brain Research Imaging Centre of Edinburgh. <http://www.bric.ed.ac.uk>. (accessed June 2016).
- [2] LST-A lesion segmentation tool for SPM. www.statistical-modelling.de/lst.html. (accessed November 2015).
- [3] A randomized trial of anticoagulants versus aspirin after cerebral ischemia of presumed arterial origin. The Stroke Prevention in Reversible Ischemia Trial (SPIRIT) Study Group. *Annals of neurology*, 42(6):857–65, 1997.
- [4] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(2):124007, 2017.
- [5] J. Adler and O. Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 2018.
- [6] F. Admiraal-Behloul, D. van den Heuvel, H. Olofsen, M. J. van Osch, J. van der Grond, M. van Buchem, and J. Reiber. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *NeuroImage*, 28(3):607–617, 2005.
- [7] H. K. Aggarwal, M. P. Mani, and M. Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE transactions on medical imaging*, 38(2):394–405, 2018.
- [8] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *ICCV*, pages 37–45, 2015.

- [9] A. Aizerman, E. M. Braverman, and L. I. Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.
- [10] M. Akçakaya, S. Moeller, S. Weingärtner, and K. Uğurbil. Scan-specific robust artificial-neural-networks for k-space interpolation (RAKI) reconstruction: Database-free deep learning for fast imaging. *Magnetic resonance in medicine*, 81(1):439–453, 2019.
- [11] R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, A. Belopolsky, Y. Bengio, A. Bergeron, J. Bergstra, V. Bisson, J. Blecher Snyder, N. Bouchard, N. Boulanger-Lewandowski, X. Bouthillier, A. de Brébisson, O. Breuleux, P.-L. Carrier, K. Cho, J. Chorowski, P. Christiano, T. Cooijmans, M.-A. Côté, M. Côté, A. Courville, Y. N. Dauphin, O. Delalleau, J. Demouth, G. Desjardins, S. Dieleman, L. Dinh, M. Ducoffe, V. Dumoulin, S. Ebrahimi Kahou, D. Erhan, Z. Fan, O. Firat, M. Germain, X. Glorot, I. Goodfellow, M. Graham, C. Gulcehre, P. Hamel, I. Harlouchet, J.-P. Heng, B. Hidasi, S. Honari, A. Jain, S. Jean, K. Jia, M. Korobov, V. Kulkarni, A. Lamb, P. Lamblin, E. Larsen, C. Laurent, S. Lee, S. Lefrancois, S. Lemieux, N. Léonard, Z. Lin, J. A. Livezey, C. Lorenz, J. Lowin, Q. Ma, P.-A. Manzagol, O. Mastropietro, R. T. McGibbon, R. Memisevic, B. van Merriënboer, V. Michalski, M. Mirza, A. Orlandi, C. Pal, R. Pascanu, M. Pezeshki, C. Raffel, D. Renshaw, M. Rocklin, A. Romero, M. Roth, P. Sadowski, J. Salvatier, F. Savard, J. Schlüter, J. Schulman, G. Schwartz, I. V. Serban, D. Serdyuk, S. Shabanian, É. Simon, S. Spieckermann, S. R. Subramanyam, J. Sygnowski, J. Tanguay, G. van Tulder, J. Turian, S. Urban, P. Vincent, F. Visin, H. de Vries, D. Warde-Farley, D. J. Webb, M. Willson, K. Xu, L. Xue, L. Yao, S. Zhang, and Y. Zhang. Theano: A Python framework for fast computation of mathematical expressions, url = <http://arxiv.org/abs/1605.02688>, volume = abs/1605.02688, year = 2016. *arXiv e-prints*, May.
- [12] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond. Probabilistic segmentation of white matter lesions in MR imaging. *NeuroImage*, 21(3):1037–1044, 2004.

- [13] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond. Probabilistic segmentation of white matter lesions in MR imaging. *NeuroImage*, 21(3):1037–1044, 2004.
- [14] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [15] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.
- [16] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006.
- [17] J. Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95 – 113, 2007.
- [18] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.
- [19] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.
- [20] W. Bai, C. Chen, G. Tarroni, J. Duan, F. Guitton, S. E. Petersen, Y. Guo, P. M. Matthews, and D. Rueckert. Self-supervised learning for cardiac MR image segmentation by anatomical position prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 541–549. Springer, 2019.
- [21] W. Bai, M. Sinclair, G. Tarroni, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 2018.

- [22] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data*, 4:170117, 2017.
- [23] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. An unsupervised learning model for deformable medical image registration. In *CVPR*, pages 9252–9260, 2018.
- [24] R. Barber, P. Scheltens, A. Gholkar, C. Ballard, I. McKeith, P. Ince, R. Perry, and J. O’Brien. White matter lesions on magnetic resonance imaging in dementia with Lewy bodies, Alzheimer’s disease, vascular dementia, and normal aging. *Journal of Neurology, Neurosurgery & Psychiatry*, 67(1):66–72, 1999.
- [25] J. J. Bartko. Measurement and reliability: statistical thinking considerations. *Schizophrenia bulletin*, 17(3):483–9, 1991.
- [26] D. Batenkov, Y. Romano, and M. Elad. On the global-local dichotomy in sparsity modeling. In *Compressed Sensing and its Applications*, pages 1–53. Springer, 2017.
- [27] C. F. Baumgartner, L. M. Koch, M. Pollefeys, and E. Konukoglu. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 111–119. Springer, 2017.
- [28] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei Fei. What’s the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016.
- [29] M. F. Beg, M. I. Miller, A. Trounev, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, 2005.
- [30] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework

- for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006.
- [31] K. Bendfeldt, J. O. Blumhagen, H. Egger, P. Loetscher, N. Denier, P. Kuster, S. Traud, N. Mueller-Lenke, Y. Naegelin, A. Gass, J. Hirsch, L. Kappos, T. E. Nichols, E.-W. Radue, and S. J. Borgwardt. Spatiotemporal distribution pattern of white matter lesion volumes and their association with regional grey matter volume reductions in relapsing-remitting multiple sclerosis. *Human brain mapping*, 31(10):1542–55, oct 2010.
- [32] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.
- [33] A. Birenbaum and H. Greenspan. Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. In *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 58–67. Springer, 2016.
- [34] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.
- [35] J. M. Bland and D. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.
- [36] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.
- [37] C. Bowles, C. Qin, C. Ledig, R. Guerrero, R. Gunn, A. Hammers, E. Sakka, D. A. Dickie, M. V. Hernández, N. Royle, et al. Pseudo-healthy image synthesis for white matter lesion segmentation. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 87–96. Springer, 2016.
- [38] C. Bowles, C. Qin, C. Ledig, R. Guerrero, A. Hammers, E. Sakka, D. A. Dickie, B. Tijms, A. W. Lemstra, and W. V. D. Flier. Pseudo-Healthy Image Synthesis for White Matter Lesion Segmentation. In *International Workshop on Simulation and Synthesis in Medical Imaging*, volume 9968, pages 87–96. Springer International Publishing, 2016.

- [39] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
- [40] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.
- [41] T. Brosch, Y. Yoo, L. Y. Tang, D. K. Li, A. Traboulsee, and R. Tam. Deep convolutional encoder networks for multiple sclerosis lesion segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 3–11. Springer, 2015.
- [42] T. Brott, J. R. Marler, C. P. Olinger, H. P. Adams, T. Tomsick, W. G. Barsan, J. Biller, R. Eberle, V. Hertzberg, and M. Walker. Measurements of acute cerebral infarction: lesion size by computed tomography. *Stroke*, 20(7), 1989.
- [43] J. Caballero, C. Ledig, A. Aitken, et al. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, 2017.
- [44] J. Caballero, A. N. Price, D. Rueckert, and J. V. Hajnal. Dictionary learning and time sparsity for dynamic MR data reconstruction. *IEEE Transactions on Medical Imaging*, 33(4):979–994, 2014.
- [45] M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, and A. Cherubini. Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review. *Neuroinformatics*, 13(3):261–276, 2015.
- [46] X. Cao, J. Yang, Y. Gao, Y. Guo, G. Wu, and D. Shen. Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis. *MedIA*, 41:18–31, 2017.
- [47] X. Cao, J. Yang, J. Zhang, D. Nie, M. Kim, Q. Wang, and D. Shen. Deformable image registration based on similarity-steered cnn regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–308. Springer, 2017.

- [48] M. J. Cardoso, C. H. Sudre, M. Modat, and S. Ourselin. Template-based multimodal joint generative model of brain data. In *Information Processing in Medical Imaging*, volume 9123, pages 17–29. Springer International Publishing, 2015.
- [49] D. T. Chard, C. M. Griffin, G. J. M. Parker, R. Kapoor, A. J. Thompson, and D. H. Miller. Brain atrophy in clinically early relapsing-remitting multiple sclerosis. *Brain: a journal of neurology*, 125(2):327–37, 2002.
- [50] L. Chen, T. Tong, C. P. Ho, R. Patel, D. Cohen, A. C. Dawson, O. Halse, O. Geraghty, P. E. Rinne, C. J. White, et al. Identification of cerebral small vessel disease using multiple instance learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 523–530. Springer, 2015.
- [51] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.
- [52] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, pages 686–695, 2017.
- [53] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [54] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.
- [55] J. Cohen. *Statistical power analysis for the behavioral sciences*, page 40. Lawrence Erlbaum Associates, 2nd edition, 1988.
- [56] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

- [57] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho. End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2017.
- [58] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018.
- [59] A. V. Dalca, R. Sridharan, L. Cloonan, K. M. Fitzpatrick, A. Kanakis, K. L. Furie, J. Rosand, O. Wu, M. Sabuncu, N. S. Rost, et al. Segmentation of cerebrovascular pathologies in stroke patients with spatial and shape priors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 17, page 773. NIH Public Access, 2014.
- [60] S. Damangir, A. Manzouri, K. Oppedal, S. Carlsson, M. J. Firbank, H. Sonnesyn, O.-B. Tysnes, J. T. O’Brien, M. K. Beyer, E. Westman, D. Aarsland, L.-O. Wahlund, and G. Spulber. Multispectral MRI segmentation of age related white matter changes using a cascade of support vector machines. *Journal of the Neurological Sciences*, 322(1):211–216, 2012.
- [61] S. Datta, G. Tao, R. He, J. S. Wolinsky, and P. A. Narayana. Improved cerebellar tissue classification on magnetic resonance images of brain. *Journal of Magnetic Resonance Imaging*, 29(5):1035–1042, 2009.
- [62] C. Davatzikos. Spatial transformation and registration of brain images using elastically deformable models. *Comput. Vis. Image Underst.*, 66(2):207–222, 1997.
- [63] A. Davy, M. Havaei, D. Warde-farley, A. Biard, L. Tran, A. Courville, H. Larochelle, C. Pal, and Y. Bengio. Brain tumor segmentation with deep neural networks. 2014.
- [64] M. De Craene, G. Piella, O. Camara, et al. Temporal diffeomorphic free-form deformation: Application to motion and strain estimation from 3D echocardiography. *Medical image analysis*, 16(2):427–450, 2012.

- [65] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.
- [66] A. Derntl, C. Plant, P. Gruber, S. Wegener, J. S. Bauer, and B. H. Menze. Stroke lesion segmentation using a probabilistic atlas of cerebral vascular territories. In *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 21–32. Springer, 2015.
- [67] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
- [68] S. Dieleman, J. Schlüter, C. Raffel, E. Olson, S. K. Sønderby, D. Nouri, D. Maturana, M. Thoma, E. Battenberg, J. Kelly, J. D. Fauw, M. Heilman, D. M. de Almeida, B. McFee, H. Weideman, G. Takács, P. de Rivaz, J. Crall, G. Sanders, K. Rasul, C. Liu, G. French, and J. Degraeve. Lasagne: First release., Aug. 2015.
- [69] S. I. Dimitriadis, D. Liparas, A. D. N. Initiative, et al. How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer’s disease: from Alzheimer’s disease neuroimaging initiative (ADNI) database. *Neural regeneration research*, 13(6):962, 2018.
- [70] C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.
- [71] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. In *ICCV*, 2017.
- [72] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.
- [73] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199. Springer, 2014.

- [74] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016.
- [75] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo. Automatic brain tumor detection and segmentation using U-net based fully convolutional networks. In *annual conference on medical image understanding and analysis*, pages 506–517. Springer, 2017.
- [76] G. Dugas-Phocion, M. A. G. Ballester, G. Malandain, C. Lebrun, and N. Ayache. Improved EM-based tissue segmentation and partial volume effect quantification in multi-sequence brain MRI. In *Medical Image Computing and Computer-Assisted Intervention*, pages 26–33. Springer, 2004.
- [77] J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.
- [78] T. Eo, Y. Jun, T. Kim, J. Jang, H.-J. Lee, and D. Hwang. KIKI-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. *Magnetic resonance in medicine*, 80(5):2188–2201, 2018.
- [79] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline. Machine learning for medical imaging. *Radiographics*, 37(2):505–515, 2017.
- [80] J. Fan, X. Cao, Z. Xue, P.-T. Yap, and D. Shen. Adversarial similarity network for evaluating image alignment in deep learning based registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 739–746. Springer, 2018.
- [81] F. Fazekas, J. B. Chawluk, A. Alavi, H. I. Hurtig, and R. A. Zimmerman. MR signal abnormalities at 1.5 T in Alzheimer’s dementia and normal aging. *AJR. American journal of roentgenology*, 149(2):351–6, 1987.
- [82] F. Forbes, S. Doyle, D. Garcia-Lorenzo, C. Barillot, and M. Dojat. Adaptive weighted fusion of multiple MR sequences for brain lesion segmentation. In *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, pages 69–72. IEEE, 2010.

- [83] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, J. Pineau, et al. An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 11(3-4):219–354, 2018.
- [84] O. Freifeld, H. Greenspan, and J. Goldberger. Multiple sclerosis lesion detection using constrained GMM and curve evolution. *Journal of Biomedical Imaging*, 2009:14, 2009.
- [85] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multi-modal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.
- [86] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical image analysis*, 17(1):1–18, 2013.
- [87] E. Geremia, O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi, and N. Ayache. Spatial decision forests for ms lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, 57(2):378–390, 2011.
- [88] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. W. van Uden, C. I. Sanchez, G. Litjens, F.-E. de Leeuw, B. van Ginneken, E. Marchiori, and B. Platel. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, 7(1):5110, 2017.
- [89] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. W. M. van Uder, F. E. de Leeuw, E. Marchiori, B. van Ginneken, and B. Platel. Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. In *International Symposium on Biomedical Imaging*, pages 1414–1417. IEEE, apr 2016.
- [90] D. Giavarina. Understanding Bland Altman analysis. *Biochemia medica*, 25(2):141–151, 2015.
- [91] E. Gibson, F. Gao, S. E. Black, and N. J. Lobaugh. Automatic segmentation of white matter hyperintensities in the elderly using FLAIR images at 3T. *Journal of Magnetic Resonance Imaging*, 31(6):1311–1322, 2010.

- [92] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.
- [93] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios. Dense image registration through MRFs and efficient linear programming. *Medical Image Analysis*, 12(6):731 – 741, 2008.
- [94] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. 2016.
- [95] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [96] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. DRAW: A recurrent neural network for image generation. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1462–1471, 2015.
- [97] L. Griffanti, G. Zamboni, A. Khan, L. Li, G. Bonifacio, V. Sundaresan, U. G. Schulz, W. Kuker, M. Battaglini, P. M. Rothwell, et al. BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *Neuroimage*, 141:191–205, 2016.
- [98] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002.
- [99] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *NIPS*, pages 5767–5777, 2017.
- [100] E. Haber and J. Modersitzki. Intensity gradient based registration and fusion of multi-modal images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 726–733. Springer, 2006.

- [101] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magnetic resonance in medicine*, 79:3055–3071, 2018.
- [102] Y. Han and J. C. Ye. k-space deep learning for accelerated MRI. *arXiv preprint arXiv:1805.03779*, 2018.
- [103] Y. Han, J. Yoo, H. H. Kim, H. J. Shin, K. Sung, and J. C. Ye. Deep learning with domain adaptation for accelerated projection-reconstruction MR. *Magnetic Resonance in Medicine*, 80(3):1189–1205, 2018.
- [104] R. Harmouche, L. Collins, D. Arnold, S. Francis, and T. Arbel. Bayesian ms lesion classification modeling regional and local spatial information. In *18th International Conference on Pattern Recognition*, volume 3, pages 984–987. IEEE, 2006.
- [105] G. Haskins, U. Kruger, and P. Yan. Deep learning in medical image registration: A survey. *arXiv preprint arXiv:1903.02026*, 2019.
- [106] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [107] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.
- [108] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, M. Brady, and J. A. Schnabel. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. *MedIA*, 16(7):1423–1435, 2012.
- [109] M. Hernandez, M. N. Bossa, and S. Olmos. Registration of anatomical images using paths of diffeomorphisms parameterized with stationary vector field flows. *International Journal of Computer Vision*, 85(3):291–306, 2009.
- [110] V. Hernández, Z. Morris, D. A. Dickie, N. A. Royle, S. Munoz Maniega, B. S. Aribisala,

- M. E. Bastin, I. J. Deary, and J. M. Wardlaw. Close correlation between quantitative and qualitative assessments of white matter lesions. *Neuroepidemiology*, 40(1):13–22, 2012.
- [111] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [112] X. Hu and T. Parrish. Reduction of field of view for dynamic imaging. *Magnetic Resonance in Medicine*, 31(6):691–694, 1994.
- [113] Y. Hu, M. Modat, E. Gibson, N. Ghavami, E. Bonmati, C. M. Moore, M. Emberton, J. A. Noble, D. C. Barratt, and T. Vercauteren. Label-driven weakly-supervised learning for multimodal deformable image registration. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1070–1074. IEEE, 2018.
- [114] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, M. Emberton, et al. Weakly-supervised convolutional neural networks for multimodal image registration. *MedIA*, 49:1–13, 2018.
- [115] K. Hua, J. Zhang, S. Wakana, H. Jiang, X. Li, D. S. Reich, P. A. Calabresi, J. J. Pekar, P. C. van Zijl, and S. Mori. Tract probability maps in stereotaxic spaces: Analyses of white matter anatomy and tract-specific quantification. *NeuroImage*, 39(1):336–347, 2008.
- [116] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [117] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.
- [118] Y. Huang, W. Wang, and L. Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Advances in Neural Information Processing Systems*, pages 235–243, 2015.

- [119] H. J. Huppertz, J. Wagner, B. Weber, P. House, and H. Urbach. Automated quantitative FLAIR analysis in hippocampal sclerosis. *Epilepsy research*, 97(1):146–156, 2011.
- [120] T. Hyun Kim, M. S. Sajjadi, M. Hirsch, and B. Scholkopf. Spatio-temporal transformer network for video restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 106–122, 2018.
- [121] V. Ithapu, V. Singh, C. Lindner, B. P. Austin, C. Hinrichs, C. M. Carlsson, B. B. Bendlin, and S. C. Johnson. Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer’s disease risk and aging studies. *Human brain mapping*, 35(8):4219–4235, 2014.
- [122] L. Itti, L. Chang, and T. Ernst. Segmentation of progressive multifocal leukoencephalopathy lesions in fluid-attenuated inversion recovery magnetic resonance imaging. *Journal of Neuroimaging*, 11(4):412–417, oct 2001.
- [123] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015.
- [124] S. Jégou, M. Drozdal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017.
- [125] M. Jenkinson, P. Bannister, M. Brady, and S. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002.
- [126] Z. Ji, J. Liu, G. Cao, Q. Sun, and Q. Chen. Robust spatially constrained fuzzy c-means algorithm for brain MR image segmentation. *Pattern Recognition*, 47(7):2454–2466, 2014.
- [127] R. Jones, O. Haraldseth, T. Müller, P. Rinck, and A. Øksendal. K-space substitution: a novel dynamic imaging technique. *Magnetic resonance in medicine*, 29(6):830–834, 1993.

- [128] H. Jung, K. Sung, K. S. Nayak, E. Y. Kim, and J. C. Ye. k-t focuss: a general compressed sensing framework for high resolution dynamic mri. *Magnetic resonance in medicine*, 61(1):103–116, 2009.
- [129] H. Jung, J. C. Ye, and E. Y. Kim. Improved k-t BLAST and k-t SENSE using FOCUSS. *Physics in medicine and biology*, 52(11):3201, 2007.
- [130] K. Kamnitsas, L. Chen, C. Ledig, D. Rueckert, and B. Glocker. Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. In *ISLES-International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.
- [131] 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*, 36:61–78, 2017.
- [132] A. Karpathy and F. F. Li. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [133] J. Kawa and E. Pietka. Kernelized fuzzy c-means method in fast segmentation of demyelination plaques in multiple sclerosis. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5616–5619. IEEE, 2007.
- [134] Y. Kawata, H. Arimura, Y. Yamashita, T. Magome, M. Ohki, F. Toyofuku, Y. Higashida, and K. Tsuchiya. Computer-aided evaluation method of white matter hyperintensities related to subcortical vascular dementia based on magnetic resonance imaging. *Computerized Medical Imaging and Graphics*, 34(5):370–376, 2010.
- [135] S. Kazemnia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay. GANs for medical image analysis. *arXiv preprint arXiv:1809.06222*, 2018.
- [136] R. Khayati, M. Vafadust, F. Towhidkhah, and M. Nabavi. Fully automatic segmentation of multiple sclerosis lesions in brain mr flair images using adaptive mixtures method and markov random field model. *Computers in biology and medicine*, 38(3):379–390, 2008.

- [137] R. Kikinis, C. R. Guttmann, D. Metcalf, W. M. Wells, G. J. Ettinger, H. L. Weiner, and F. A. Jolesz. Quantitative follow-up of patients with multiple sclerosis using MRI: technical aspects. *Journal of Magnetic Resonance Imaging*, 9(4):519–530, 1999.
- [138] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations*, 2015.
- [139] S. Klein, M. Staring, and J. P. Pluim. Evaluation of optimization methods for nonrigid medical image registration using mutual information and b-splines. *IEEE Transactions on Medical Imaging*, 16(12):2879–2890, 2007.
- [140] F. Knoll, K. Hammernik, C. Zhang, S. Moeller, T. Pock, D. K. Sodickson, and M. Akcakaya. Deep learning methods for parallel magnetic resonance image reconstruction. *arXiv preprint arXiv:1904.01112*, 2019.
- [141] J. Krebs, T. Mansi, H. Delingette, L. Zhang, F. C. Ghesu, S. Miao, A. K. Maier, N. Ayache, R. Liao, and A. Kamen. Robust non-rigid registration through agent-based action learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 344–352. Springer, 2017.
- [142] J. Krebs, T. Mansi, B. Mailhé, N. Ayache, and H. Delingette. Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 101–109. Springer, 2018.
- [143] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [144] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3677, 2016.
- [145] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

- [146] Z. Lao, D. Shen, D. Liu, A. F. Jawad, E. R. Melhem, L. J. Launer, R. N. Bryan, and C. Davatzikos. Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. *Academic radiology*, 15(3):300–313, 2008.
- [147] P. C. Lauterbur et al. Image formation by induced local interactions: examples employing nuclear magnetic resonance.
- [148] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.
- [149] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [150] C. Ledig, R. A. Heckemann, A. Hammers, J. C. Lopez, V. F. Newcombe, A. Makropoulos, J. Lötjönen, D. K. Menon, and D. Rueckert. Robust whole-brain segmentation: application to traumatic brain injury. *Medical image analysis*, 21(1):40–58, 2015.
- [151] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.
- [152] D. Lee, J. Yoo, and J. C. Ye. Deep artifact learning for compressed sensing and parallel MRI. *arXiv preprint arXiv:1703.01120*, 2017.
- [153] H. Y. Lee, H. Y. Tseng, J. B. Huang, M. Singh, and M. H. Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision*, 2018.
- [154] M. Liang and X. Hu. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3367–3375, 2015.
- [155] S. G. Lingala, Y. Hu, E. Dibella, and M. Jacob. Accelerated dynamic MRI exploiting sparsity and low-rank structure: K-t SLR. *IEEE Transactions on Medical Imaging*, 30(5):1042–1054, 2011.

- [156] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- [157] J. Liu, C. D. Smith, and H. Chebrolu. Automatic multiple sclerosis detection based on integrated square estimation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 31–38. IEEE, 2009.
- [158] M. Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.
- [159] W. Liu, G. Hua, and J. Smith. Unsupervised one-class learning for automatic outlier removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3826–3833, 2014.
- [160] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.
- [161] W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.
- [162] K.-O. Löuvbld, A. E. Baird, G. Schlaug, A. Benfield, B. Siewert, B. Voetsch, A. Connor, C. Burzynski, R. R. Edelman, and S. Warach. Ischemic lesion volumes in acute stroke by diffusion-weighted magnetic resonance imaging correlate with clinical outcome. *Annals of Neurology*, 42(2):164–170, aug 1997.
- [163] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. Compressed sensing MRI. *IEEE signal processing magazine*, 25(2):72–82, 2008.
- [164] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

- [165] D. Mahapatra, B. Antony, S. Sedai, and R. Garnavi. Deformable medical image registration using generative adversarial networks. In *International Symposium on Biomedical Imaging*, pages 1449–1453. IEEE, 2018.
- [166] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, et al. ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Medical image analysis*, 35:250–269, 2017.
- [167] O. Maier, C. Schröder, N. D. Forkert, T. Martinetz, and H. Handels. Classifiers for ischemic stroke lesion segmentation: a comparison study. *PloS one*, 10(12):e0145118, 2015.
- [168] O. Maier, M. Wilms, J. von der Gablentz, U. Krämer, and H. Handels. Ischemic stroke lesion segmentation in multi-spectral MR images with support vector machine classifiers. In *SPIE Medical Imaging*, pages 903504–903504. International Society for Optics and Photonics, 2014.
- [169] O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, T. F. Münte, and H. Handels. Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences. *Journal of neuroscience methods*, 240:89–100, 2015.
- [170] P. Maillard, N. Delcroix, F. Crivello, C. Dufouil, S. Gicquel, M. Joliot, N. Tzourio-Mazoyer, A. Alperovitch, C. Tzourio, and B. Mazoyer. An automated procedure for the assessment of white matter hyperintensities by multispectral (T1, T2, PD) MRI and an evaluation of its between-centre reproducibility based on two large community databases. *Neuroradiology*, 50(1):31–42, 2008.
- [171] J. A. Maintz, P. A. van den Elsen, and M. A. Viergever. 3D multimodality medical image registration using morphological tools. *Image and vision computing*, 19(1-2):53–62, 2001.
- [172] M. E. Maron. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, 8(3):404–417, 1961.

- [173] K. Marstal, F. Berendsen, M. Staring, and S. Klein. SimpleElastix: A user-friendly, multi-lingual library for medical image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.
- [174] P. McClure, F. Khalifa, A. Soliman, M. A. El-Ghar, G. Gimelfarb, A. Elmagraby, and A. El-Baz. A novel NMF guided level-set for DWI prostate segmentation. *J Comput Sci Syst Biol*, 7(6):209–216, 2014.
- [175] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 5(4):115–133, 1943.
- [176] R. McKinley, T. Gundersen, F. Wagner, A. Chan, R. Wiest, and M. Reyes. Nabla-net: a deep dag-like convolutional architecture for biomedical image segmentation: application to white-matter lesion segmentation in multiple sclerosis. *MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, page 37, 2016.
- [177] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *TMI*, 34(10):1993, 2015.
- [178] X. Miao, S. G. Lingala, Y. Guo, T. Jao, M. Usman, C. Prieto, and K. S. Nayak. Accelerated cardiac cine MRI using locally low rank and finite difference constraints. *Magnetic resonance imaging*, 34(6):707–714, 2016.
- [179] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [180] A. Mortazi, R. Karim, K. Rhode, J. Burt, and U. Bagci. CardiacNET: segmentation of left atrium and proximal pulmonary veins from MRI using multi-view CNN. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 377–385. Springer, 2017.
- [181] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines.

- In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.
- [182] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 417–425. Springer, 2017.
- [183] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.
- [184] R. Otazo, E. Candès, and D. K. Sodickson. Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components. *Magnetic Resonance in Medicine*, 73(3):1125–1136, 2015.
- [185] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. *Medical image analysis*, 15(4):622–639, 2011.
- [186] J. Pallant. *SPSS Survival Manual: A Step By Step Guide to Data Analysis Using SPSS*. 4th edition edition, 2010.
- [187] V. Patraucean, A. Handa, and R. Cipolla. Spatio-temporal video autoencoder with differentiable memory. *The International Conference on Learning Representations workshop*, 2016.
- [188] J. Patravali, S. Jain, and S. Chilamkurthy. 2D-3D fully convolutional neural networks for cardiac MR segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 130–139. Springer, 2017.
- [189] X. Pennec, P. Cachier, and N. Ayache. Understanding the demon’s algorithm: 3d non-rigid registration by gradient descent. In *Medical Image Computing and Computer-Assisted Intervention*, volume 26, pages 597–606, 1999.

- [190] D. Perperidis, R. H. Mohiaddin, and D. Rueckert. Spatio-temporal free-form registration of cardiac mr image sequences. *Medical image analysis*, 9(5):441–456, 2005.
- [191] J. P. Pluim, J. A. Maintz, and M. A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging*, 22(8):986–1004, 2003.
- [192] C. H. Polman, S. C. Reingold, G. Edan, M. Filippi, H.-P. Hartung, L. Kappos, F. D. Lublin, L. M. Metz, H. F. McFarland, P. W. O’Connor, et al. Diagnostic criteria for multiple sclerosis: 2005 revisions to the “McDonald Criteria”. *Annals of neurology*, 58(6):840–846, 2005.
- [193] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger. SENSE: sensitivity encoding for fast MRI. *Magnetic resonance in medicine*, 42(5):952–962, 1999.
- [194] C. Qin, W. Bai, J. Schlemper, S. E. Petersen, S. K. Piechnik, S. Neubauer, and D. Rueckert. Joint learning of motion estimation and segmentation for cardiac MR image sequences. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018.
- [195] C. Qin, R. Guerrero, C. Bowles, C. Ledig, P. Scheltens, F. Barkhof, H. Rhodius-Meester, B. Tijms, A. W. Lemstra, W. M. van der Flier, B. Glocker, and D. Rueckert. A Semi-supervised Large Margin Algorithm for White Matter Hyperintensity Segmentation. In *Machine Learning in Medical Imaging*, volume 9352, pages 104–112. Springer International Publishing, 2016.
- [196] C. Qin, R. Guerrero-Moreno, C. Bowles, C. Ledig, P. Scheltens, F. Barkhof, H. Rhodius-Meester, B. Tijms, A. W. Lemstra, W. M. van der Flier, et al. A semi-supervised large margin algorithm for white matter hyperintensity segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 104–112. Springer, 2016.
- [197] C. Qin, J. Schlemper, J. Caballero, A. N. Price, J. V. Hajnal, and D. Rueckert. Convolutional recurrent neural networks for dynamic MR image reconstruction. *IEEE transactions on medical imaging*, 38(1):280–290, 2019.

- [198] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [199] V. Rajinikanth, S. C. Satapathy, S. L. Fernandes, and S. Nachiappan. Entropy based segmentation of tumor from brain MR images—A study with teaching learning based optimization. *Pattern Recognition Letters*, 2017.
- [200] S. Ravishankar and Y. Bresler. MR image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE transactions on medical imaging*, 30(5):1028–1041, 2011.
- [201] E. A. Regan, J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, D. Curran-Everett, E. K. Silverman, and J. D. Crapo. Genetic epidemiology of COPD (COPDGene) study design. *Journal of COPD*, 7(1):32–43, 2011.
- [202] I. Rekik, S. Allasonnière, T. K. Carpenter, and J. M. Wardlaw. Medical image analysis methods in mr/ct-imaged acute-subacute ischemic stroke lesion: Segmentation, prediction and insights into dynamic evolution simulation models. a critical appraisal. *NeuroImage: Clinical*, 1(1):164–178, 2012.
- [203] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, and H. Zha. Unsupervised deep learning for optical flow estimation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [204] N. Richard, M. Dojat, and C. Garbay. Distributed Markovian segmentation: Application to MR brain scans. *Pattern Recognition*, 40(12):3467–3480, 2007.
- [205] A. Roche, X. Pennec, G. Malandain, and N. Ayache. Rigid registration of 3-D ultrasound with MR images: a new approach combining intensity and gradient information. *TMI*, 20(10):1038–1049, 2001.
- [206] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

- [207] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [208] A. K. Rudra, M. Sen, A. S. Chowdhury, A. Elnakib, and A. El-Baz. 3D graph cut with new edge weights for cerebral white matter segmentation. *Pattern Recognition Letters*, 32(7):941–947, 2011.
- [209] D. Rueckert, L. I. Sonoda, C. Hayes, et al. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999.
- [210] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017.
- [211] B. R. Sajja, S. Datta, R. He, M. Mehta, R. K. Gupta, J. S. Wolinsky, and P. A. Narayana. Unified approach for multiple sclerosis lesion segmentation on brain mri. *Annals of biomedical engineering*, 34(1):142–151, 2006.
- [212] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.
- [213] J. N. Sarvaiya, S. Patnaik, and S. Bombaywala. Image registration by template matching using normalized cross-correlation. In *International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pages 819–822. IEEE, 2009.
- [214] J. Schlemper, J. Caballero, J. V. Hajnal, A. Price, and D. Rueckert. A deep cascade of convolutional neural networks for MR image reconstruction. In *International Conference on Information Processing in Medical Imaging*, pages 647–658, 2017.
- [215] J. Schlemper, J. Caballero, J. V. Hajnal, A. Price, and D. Rueckert. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Transactions on Medical Imaging*, 37(2), 2018.

- [216] J. Schlemper, O. Oktay, W. Bai, et al. Cardiac MR segmentation from undersampled k-space using deep latent representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018.
- [217] P. Schmidt. *Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging*. PhD thesis, Ludwig-Maximilians-Universität München, 2017.
- [218] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förchler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer, B. Hemmer, and M. Mühlau. An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. *NeuroImage*, 59(4):3774–3783, 2012.
- [219] J. A. Schnabel, D. Rueckert, M. Quist, J. M. Blackall, A. D. Castellano-Smith, T. Hartkens, G. P. Penney, W. A. Hall, H. Liu, C. L. Truwit, et al. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In *Medical Image Computing and Computer-Assisted Intervention*, pages 573–581. Springer, Heidelberg, 2001.
- [220] M. Seitzer, G. Yang, J. Schlemper, O. Oktay, T. Würfl, V. Christlein, T. Wong, R. Mohiaddin, D. Firmin, J. Keegan, et al. Adversarial and perceptual refinement for compressed sensing mri reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–240. Springer, 2018.
- [221] C. E. Shannon. Communication in the presence of noise. *Proceedings of the IEEE*, 86(2):447–457, 1998.
- [222] D. Shen and C. Davatzikos. HAMMER: hierarchical attribute matching mechanism for elastic registration. In *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 29–36. IEEE, 2001.
- [223] D. Shen, H. Sundar, Z. Xue, Y. Fan, and H. Litt. Consistent estimation of cardiac motions by 4D image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 902–910, 2005.

- [224] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian. Multi-scale convolutional neural networks for lung nodule classification. In *International Conference on Information Processing in Medical Imaging*, pages 588–599. Springer, 2015.
- [225] W. Shi, X. Zhuang, H. Wang, et al. A comprehensive cardiac motion estimation framework using both untagged and 3-D tagged MR images based on nonrigid registration. *IEEE transactions on medical imaging*, 31(6):1263–1275, 2012.
- [226] N. Shiee, P.-L. Bazin, A. Ozturk, D. S. Reich, P. A. Calabresi, and D. L. Pham. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, 49(2):1524–1535, 2010.
- [227] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, et al. A deep metric for multimodal registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 10–18, 2016.
- [228] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, pages 1–14, sep 2015.
- [229] S. Singh, A. Srivastava, L. Mi, R. J. Caselli, K. Chen, D. Goradia, E. M. Reiman, and Y. Wang. Deep-learning-based classification of FDG-PET data for Alzheimer’s disease categories. In *13th International Conference on Medical Information Processing and Analysis*, volume 10572, page 105720J. International Society for Optics and Photonics, 2017.
- [230] H. Sokootti, B. de Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–239. Springer, 2017.
- [231] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7):1153–1190, 2013.

- [232] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using LSTMs. In *International conference on machine learning*, pages 843–852, 2015.
- [233] C. Studholme, D. L. Hill, and D. J. Hawkes. An overlap invariant entropy measure of 3D medical image alignment. *Pattern recognition*, 32(1):71–86, 1999.
- [234] M. Styner, J. Lee, B. Chin, M. Chin, O. Commowick, H. Tran, S. Markovic-Plese, V. Jewells, and S. Warfield. 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. *MIDAS Journal*, 2008:1–6, 2008.
- [235] J. Sun, H. Li, Z. Xu, et al. Deep ADMM-Net for compressive sensing MRI. In *Advances in Neural Information Processing Systems*, pages 10–18, 2016.
- [236] Y. Sun, A. Moelker, W. J. Niessen, and T. van Walsum. Towards robust CT-ultrasound registration using deep learning methods. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 43–51. Springer, 2018.
- [237] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [238] C. Tanner, F. Ozdemir, R. Profanter, V. Vishnevsky, E. Konukoglu, and O. Goksel. Generative adversarial networks for MR-CT deformable image registration. *arXiv preprint arXiv:1807.07349*, 2018.
- [239] J.-P. Thirion. Image matching as a diffusion process: an analogy with Maxwell’s demons. *Medical Image Analysis*, 2(3):243–260, 1998.
- [240] C. Tobon-Gomez, M. De Craene, K. Mcleod, et al. Benchmarking framework for myocardial tracking and deformation algorithms: An open access database. *Medical image analysis*, 17(6):632–648, 2013.
- [241] Q. Tong, M. Ning, W. Si, X. Liao, and J. Qin. 3D deeply-supervised U-net based whole

- heart segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 224–232. Springer, 2017.
- [242] A. Trounevé. Diffeomorphisms groups and pattern matching in image analysis. *International journal of computer vision*, 28(3):213–221, 1998.
- [243] Y. H. Tsai, M. H. Yang, and M. J. Black. Video segmentation via object flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3899–3908, 2016.
- [244] J. Tsao, P. Boesiger, and K. P. Pruessmann. k-t BLAST and k-t SENSE: Dynamic MRI with high frame rate exploiting spatiotemporal correlations. *Magnetic Resonance in Medicine*, 50(5):1031–1042, 2003.
- [245] H. Uzunova, M. Wilms, H. Handels, et al. Training CNNs for image registration from few samples with model-based data augmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017.
- [246] S. Vaidya, A. Chunduru, R. Muthuganapathy, and G. Krishnamurthi. Longitudinal multiple sclerosis lesion segmentation using 3D convolutional neural networks. *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.
- [247] M. d. C. Valdés Hernández, P. A. Armitage, M. J. Thrippleton, F. Chappell, E. Sandeman, S. Muñoz Maniega, K. Shuler, and J. M. Wardlaw. Rationale, design and methodology of the image analysis protocol for studies of patients with cerebral small vessel disease and mild stroke. *Brain and Behavior*, 5(12):1–8, 2015.
- [248] M. d. C. Valdés Hernández, L. C. Maconick, S. Muñoz Maniega, X. Wang, S. Wiseman, P. A. Armitage, F. N. Doubal, S. Makin, C. L. M. Sudlow, M. S. Dennis, I. J. Deary, M. Bastin, and J. M. Wardlaw. A comparison of location of acute symptomatic vs. ‘silent’ small vessel lesions. *International Journal of Stroke*, 10(7):1044–1050, 2015.
- [249] S. Valverde, M. Cabezas, E. Roura, S. González-Villa, J. Salvi, A. Oliver, and X. Lladó. Multiple sclerosis lesion detection and segmentation using a convolutional neural network

- of 3D patches. *MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, page 75, 2016.
- [250] J. van Gijn. Leukoaraiosis and vascular dementia. *Neurology*, 51(3 Suppl 3):S3–8, 1998.
- [251] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Transactions on Medical Imaging*, 20(8):677–688, 2001.
- [252] H. Van Nguyen, K. Zhou, and R. Vemulapalli. Cross-domain synthesis of medical images using efficient location-sensitive deep network. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9349, pages 677–684. Springer International Publishing, 2015.
- [253] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.
- [254] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Non-parametric diffeomorphic image registration with the demons algorithm. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2007.
- [255] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
- [256] P. Viola and W. M. Wells III. Alignment by maximization of mutual information. *IJCV*, 24(2):137–154, 1997.
- [257] M. T. Vlaardingerbroek and J. A. Boer. *Magnetic resonance imaging: theory and practice*. Springer Science & Business Media, 2013.
- [258] H. Wang and D.-Y. Yeung. Towards bayesian deep learning: A survey. *arXiv preprint arXiv:1604.01662*, 2016.

- [259] S. Wang, N. Huang, T. Zhao, Y. Yang, L. Ying, and D. Liang. 1D Partial Fourier Parallel MR imaging with deep convolutional neural network. In *ISMRM 25th Annual Meeting and Exhibition*, volume 47, pages 2016–2017, 2017.
- [260] S. Wang, Z. Su, L. Ying, X. Peng, and D. Liang. Exploiting deep convolutional neural network for fast magnetic resonance imaging. In *ISMRM 24th Annual Meeting and Exhibition*, 2016.
- [261] X. Wang, M. C. Valdés Hernández, F. Doubal, F. M. Chappell, and J. M. Wardlaw. How much do focal infarcts distort white matter lesions and global cerebral atrophy measures?. *Cerebrovascular diseases (Basel, Switzerland)*, 34(5-6):336–42, 2012.
- [262] Y. Wang, J. A. Catindig, S. Hilal, H. W. Soon, E. Ting, T. Y. Wong, N. Venketasubramanian, C. Chen, and A. Qiu. Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts. *Neuroimage*, 60(4):2379–2388, 2012.
- [263] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [264] J. M. Wardlaw, M. C. V. Hernández, and S. Muñoz-Maniega. What are white matter hyperintensities made of? Relevance to vascular cognitive impairment. *Journal of the American Heart Association*, 4(6):e001140, 2015.
- [265] J. M. Wardlaw and L. Pantoni. *Sporadic small vessel disease: pathogenic aspects*, pages 52–63. Cambridge University Press, Cambridge, 2014.
- [266] J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T. O’Brien, F. Barkhof, O. R. Benavente, S. E. Black, C. Brayne, M. Breteler, H. Chabriat, C. DeCarli, F.-E. de Leeuw, F. Doubal, M. Duering, N. C. Fox, S. Greenberg, V. Hachinski, I. Kilimann, V. Mok, R. van Oostenbrugge, L. Pantoni, O. Speck, B. C. M. Stephan, S. Teipel, A. Viswanathan, D. Werring, C. Chen, C. Smith, M. van Buchem, B. Norrving, P. B. Gorelick, and M. Dichgans. Neuroimaging standards for research into

- small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12(8):822 – 838, 2013.
- [267] W. Wein, S. Brunke, A. Khamene, M. R. Callstrom, and N. Navab. Automatic CT-ultrasound registration for diagnostic imaging and image-guided intervention. *Medical Image Analysis*, 12(5):577 – 585, 2008.
- [268] N. Weiss, D. Rueckert, and A. Rao. Multiple sclerosis lesion segmentation using dictionary learning and sparse coding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 8149, pages 735–742, 2013.
- [269] W. M. Wells III, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz. Adaptive segmentation of mri data. *IEEE Transactions on Medical Imaging*, 15(4):429–442, 1996.
- [270] P. Werbos. Beyond regression:” new tools for prediction and analysis in the behavioral sciences. *Ph. D. dissertation, Harvard University*, 1974.
- [271] M. Wu and J. Ye. A small sphere and large margin approach for novelty detection using training data with outliers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2088–2092, 2009.
- [272] Z. Wu, C. Shen, and A. v. d. Hengel. High-performance semantic segmentation using very deep fully convolutional networks. *arXiv preprint arXiv:1604.04339*, 2016.
- [273] S. Xingjian, Z. Chen, H. Wang, D. Y. Yeung, W.-K. Wong, and W. C. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.
- [274] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [275] M. Yaman and S. Kalkan. An iterative adaptive multi-modal stereo-vision method using mutual information. *Journal of Visual Communication and Image Representation*, 26:115–131, 2015.

- [276] F. Yang, Z. Y. Shan, and F. Kruggel. White matter lesion segmentation based on feature joint occurrence probability and χ^2 random field theory from magnetic resonance (MR) images. *Pattern Recognition Letters*, 31(9):781–790, 2010.
- [277] D. H. Ye, D. Zikic, B. Glocker, A. Criminisi, and E. Konukoglu. Modality propagation: Coherent synthesis of subject-specific scans with data-driven regularization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 8149, pages 606–613. Springer Berlin Heidelberg, 2013.
- [278] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
- [279] H. Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004.
- [280] P. Zhang et al. Multi-channel generative adversarial network for parallel magnetic resonance image reconstruction in k-space. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 180–188, 2018.
- [281] S. Zhang, K. T. Block, and J. Frahm. Magnetic resonance imaging in real time: advances using radial FLASH. *Magn. Reson. Med.*, 31(1):101–109, 2010.
- [282] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks.
- [283] D. Zikic, Y. Ioannou, M. Brown, A. Criminisi, and A. Criminisi. Segmentation of brain tumor tissues with convolutional neural networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention-BRATS*, pages 36–39, 2014.
- [284] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000, 2003.