Imperial College London
Department of Computing

Medical image synthesis in the diagnosis and study of neurodegenerative diseases

Christopher Bowles

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of Imperial College London
September 2018

Abstract

Medical imaging is a cornerstone of modern healthcare. The ability to acquire images from inside a patient has revolutionised the way doctors diagnose and treat diseases, with almost all clinical pipelines now involving imaging to some degree. The development of these imaging methods has led to the field of medical image computing, where a multitude of tools and techniques have been proposed to aid clinicians and researchers in interpreting and analysing these images. One such family of techniques involves generating synthetic medical images. Image synthesis techniques are wide and varied, ranging from basic phantoms, to disease atlases, to high resolution photo-realistic subject-specific images. Their applications are similarly diverse: for developing novel acquisition protocols, training and testing algorithms, visualising changes in disease, predicting particular image types from others, and improving image quality and resolution. This thesis examines the use of medical image synthesis with a particular focus on applications in neurodegenerative diseases. A method for synthesising subject-specific non-pathological images from pathological images is first proposed and used for the unsupervised brain lesion segmentation. We next show how generative adversarial networks (GANs) can be used to both analyse the structural changes seen in patients with Alzheimer's disease, and to add or remove these changes from patient images to produce a subject-specific prediction of disease progression. Finally we investigate how GAN-derived synthetic data can be used to increase the size of training datasets, and under what conditions this additional data can lead to an improvement across a variety of segmentation tasks. Within this context we explore two situations: where only a small amount of labelled data is available, and where a large amount of unlabelled data is also available. We show that the proposed methods can lead to significant improvements in segmentation results, especially when a small amount of labelled data is available.

Acknowledgements

People told me doing a PhD and writing a thesis would be hard, but it turns out that when you're surrounded by so many intelligent, helpful and kind people, it's actually not that bad.

First, I would of course like to thank the EPSRC, without their generous funding I'd have starved to death long ago, and everyone at the Medical Imaging CDT, for running such a brilliant programme. Next, my supervisors, Daniel Rueckert, who was always available to offer me useful advice and guidance whenever I needed it, and Alexander Hammers and Roger Gunn, who kept me on track throughout the last 3 years with their specialist knowledge and much valued feedback. I would also like to thank all my collaborators in Edinburgh, without their generosity in sharing their data, much of this work could not have been done. In particular Maria, who answered my many questions and who would always give me valuable feedback on my work. Finally I would like to thank everyone in the BiomedIA group, past and present, for providing such a stimulating environment in which to work.

On a personal side, I would like to thank all my family and friends, especially my parents for providing me a place to stay and diverting trips to the football. My friends at the CDT - Camila, Sam and Patrick, for answering my many questions about MR physics, PET reconstruction and owls. And Poppy, Alfie and Storm, for always making themselves available to play with a ball or go for a walk.

Finally I want to thank Evelyne, for keeping me company on my late night walks home, for keeping me entertained whenever I needed a break, and for humouring me as I droned on and on about fake brains. You met me for lunch on my first day, and you will meet me for dinner after my last. Thank you for everything.

Statement of Originality

I declare that the content of this thesis is my own original work, except where I have acknowledge and cited the original source. The details of permissions sought for use for copyrighted material can be found in Appendix D.

Copyright Declaration

Contents

List of Tables

List of Figures

Acronyms

2D 2-dimensional. xxix–xxxi, 23, 93, 147, 169, 173, 239, 284, 285, 287, 288, 290–296, 299, 301

3D 3-dimensional. xxx, xxxi, 23, 25, 84, 93, 119, 121, 172, 173, 178, 193, 209, 210, 234, 235, 238, 239, 284, 285, 287, 290, 292–295, 299, 300

4D 4-dimensional. 84

Aβ_{42} amyloid-β 1-42. 29, 34, 42

AD Alzheimer's Disease. xiii, xviii, xxv, xxvii, xxviii, 17, 28–36, 41–43, 45, 48, 169, 171, 175, 178, 180, 184, 185, 187, 192–198, 200–202, 204–210, 212, 218–221, 223, 226, 228–230, 233, 238

ADNI Alzheimers Disease Neuroimaging Initiative. 20, 41–43, 46, 212, 290, 292

ADNI-2 Alzheimers Disease Neuroimaging Initiative - 2. 41, 290

ADNI-3 Alzheimers Disease Neuroimaging Initiative - 3. 41

ADNI-GO Alzheimers Disease Neuroimaging Initiative - Grand Opportunity. 41

APOE Apolipoprotein E. 219, 226, 231–233

ASSD Average Symmetric Surface Distance. 87, 121, 127

AUC Area Under the Curve. xv, 45, 177, 188, 192

bvFTD Behavioural Variant Frontotemporal Dementia. 30, 35, 45

CDR Clinical Dementia Rating. xv, xxvi, 169, 175, 178, 187, 190–192

CN Normal Controls. xxvii, xxviii, 200, 201, 204–206, 208, 219–221, 224, 225, 229

CNN Convolutional Neural Network. 51, 75, 98, 144, 147, 156, 283

CRF Conditional Random Field. 102, 104, 119, 122, 123

CSF Cerebrospinal Fluid. xiv, xix, xxi, xxii, 20, 21, 24, 34, 38, 42, 45, 83, 90, 91, 109, 110, 116, 143, 146, 148–150, 155, 156

CT Computed Tomography. xviii, 8, 13, 21, 23, 25, 26, 50, 62, 69, 72, 73, 75, 76, 81, 84, 96, 97, 103, 143, 145–147

CV Coefficient of Variation. 89

DCGAN Deep Convolutional Generative Adversarial Network. 54, 98, 161, 294

DLB Dementia with Lewy bodies. 28, 32, 35, 45

DSC Dice Similarity Coefficient. xiii, xiv, xxii, xxv, xxvi, 86, 87, 89, 96, 121–123, 127, 129–131, 145, 146, 149, 150, 155, 157, 161, 173, 175, 178, 179, 183–185, 187–190, 192, 193

DSI Disease State Index. 45

DTI Diffusion Tensor Imaging. 46, 71

DWI Diffusion Weighted Imaging. 23, 24, 39

EM Expectation Maximisation. 50, 51, 69, 90–92, 112

EMCI Early Mild Cognitive Impairment. xxviii, 219–221, 229

FAD Familial Alzheimer's Disease. 29, 33, 37

FDG fluorodeoxyglucose. 26, 43, 46

FFD Free Form Deformation. xx, 105, 111

FLAIR Fluid-attenuated Inversion Recovery. xviii–xx, 24, 37–39, 41, 47, 64, 65, 71, 81, 82, 84, 85, 90, 101–107, 109–111, 113, 115–118, 122, 125, 132, 134, 135, 143, 146, 236

fMRI Functional Magnetic Resonance Imaging. 24, 46

FTD Frontotemporal Dementia. 28, 30, 31, 35, 45

GAN Generative Adversarial Network. xiv, xxi, xxii, xxiv–xxvi, xxix, xxxi, 13, 15, 17, 19, 51–57, 59, 60, 74–77, 95–98, 138, 141–148, 150, 151, 155–161, 165, 166, 169–171, 174, 175, 177, 179, 182, 192, 193, 195–203, 209–211, 216, 217, 227, 232–239, 283–285, 287, 288, 290, 298–301

GM Grey Matter. xix–xxi, 20, 21, 39, 80, 81, 83, 90–92, 105, 106, 109, 110, 113, 116–119, 125–127, 172

GMM Gaussian Mixture Model. xx, 102, 104, 107, 112, 113, 117

HD Hausdorff Distance. 87, 121, 127

IA Image Analogies. 61, 62, 64, 66, 67, 69, 73

ICC Intra Class Correlation. 88, 121, 123, 127

JS Jaccard Similarity. 86

LMCI Late Mild Cognitive Impairment. xxviii, 219–221, 226, 228–230

LNCC Local Normalised Cross-correlation. 72

LPA Logopenic Aphasia. 31

LST Lesion Segmentation Toolbox. 102, 103, 120, 122, 125, 131

LST-LGA Lesion Growth Algorithm. xxi, 90, 102, 120–123, 127–134

LST-LPA Lesion Prediction Algorithm. xiii, xiv, xxi, 90, 102, 103, 120–123, 126–134, 236

MALPEM Multi-Atlas-Label Propagation with Expectation-Maximisation based refinement. xv, xxv, xxvi, 43, 92, 105, 178, 187–192

MCI Mild Cognitive Impairment. xxvii, 200, 201, 204–206, 208

MI Mutual Information. 73

ML Machine Learning. 9

MND Motor-neuron Disease. 31

MNI Montreal Neurological Institute. 92, 105, 107, 112, 201

MR Magnetic Resonance. xviii, xxii, xxiii, xxxi, 11, 15, 17, 19, 23–26, 36–38, 41–43, 45, 46, 48, 50, 51, 62, 65, 69, 72, 75–77, 81–84, 92, 94, 96, 97, 101, 103, 132, 135, 143, 146, 151, 163, 164, 167, 169–172, 195, 209, 210, 233, 298, 299

MRA Magnetic Resonance Angiography. 24

MRI Magnetic Resonance Imaging. 8, 11, 13, 21, 23, 25, 34, 36, 39, 41, 45, 47, 62, 69, 77, 195

MS Multiple Sclerosis. 39, 41, 85, 86, 90, 102, 103, 125, 127

MS-SSIM Multi-scale structural similarity. 57, 60

MSE Mean Squared Error. 71, 78

NCC Normalised Cross-correlation. 80, 81

OASIS Open Access Series of Image Studies. 47, 48

PCA Posterior Cortical Atrophy. 29, 30

PDD Parkinson's Disease with Dementia. 32, 35

PET Positron Emission Tomography. xviii, 8, 21, 23, 26, 27, 34, 41–43, 46, 50, 62, 72, 76

PGGAN Progressive Growing of GANs. xxii, 56, 144, 145, 147, 161, 162, 164, 166, 211, 212, 216, 217, 233, 234, 238, 239

PNFA Progressive Non-Fluent Aphasia. 31

RF Radio-Frequency. 23

ROI Region of Interest. 25, 45, 76

RPC Reproducibility Coefficient. 89

RSSI Recent Small Subcortical Infarct. 39

SD Semantic Dementia. 31

SNR Signal to Noise Ratio. 73, 74, 78

SPECT Single Photon Emission Computed Tomography. 8, 21, 23, 27, 45

SSD Sum of Squared Differences. 105

SSE Sum of Squared Errors. 88

SVD Small Vessel Disease. xviii, 32, 36, 38–40, 46, 101–103, 120, 134

SVM Support Vector Machine. 118, 119

SWD Sliced Wasserstein Distance. 58, 60

tSNE t-Distributed Stochastic Neighbour Embedding. xxiv, 180, 181

UQI Universal Quality Index. 71, 78

VAE Variational Autoencoder. 52

VD Vascular Dementia. 28, 32, 33, 35, 36, 38, 45

WGAN Wasserstein Generative Adversarial Network. xxvii, 55, 197, 201–203, 205, 210, 211, 216, 233, 293, 294

WM White Matter. xix, xx, xxiii, 20, 21, 38, 39, 80–83, 90–92, 102, 103, 105, 106, 109, 110, 113, 115–119, 122, 125, 132, 135, 170, 171, 291

WMH White Matter Hyperintensity. xiv, xviii, xx, xxi, 39–41, 47, 84, 89, 93, 101, 115, 116, 120, 121, 125, 131, 143, 146, 148, 150, 155, 156

WMH$_{\text{pvo}}$ White Matter Hyperintensity of Presumed Vascular Origin. 39, 47, 104, 119, 120, 131, 143

Chapter 1

Introduction

Ever since the first images from inside the human body were taken using X-Rays in 1895, the field of medical imaging has progressed at a considerable rate. While traditional X-Ray imaging has stood the test of time and is still used today, it has been joined by ultrasound, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Single Photon Emission Computed Tomography (SPECT), among others. Each of these imaging modalities fills an important and often complementary niche in clinical practice providing greater insight into the human body in both health and disease than would have ever been possible without them.

The desire for detailed images of the brain has driven much of this progress, leading to the development of the field of neuroimaging. The modern-day clinician now has an arsenal of techniques at their disposal allowing for highly detailed images of individual brain structures, as well as precise measures of brain activity and processes such as metabolism and the accumulation of proteins.

As imaging equipment becomes more and more prevalent, so too does the demand for computational solutions to process and analyse the complex images being produced increase. As such, the field of medical image computing has grown in parallel with that of medical imaging, and now has many journals and conferences dedicated to its advancement. The overriding goal is to develop computing techniques which can leverage the acquired imaging data to extract

the maximum amount of useful information to improve patient outcomes. In pursuance of this, Machine Learning (ML) algorithms have become ubiquitous with applications all across the medical imaging spectrum: from identifying regions of interest (segmentation), to categorising whole images (classification), to deriving characteristics from images (feature extraction), to aligning multiple images (registration), to creating images from the raw data provided by the scanner (reconstruction).

Though exceptions exist, the role of ML algorithms within these applications tend to remain the same: to statistically analyse a large amount of data (to train), before applying this learning to individual cases (to test or to deploy).

The impact of medical image computing cannot be underestimated. Computers have long proven themselves adept at automating tasks which humans find laborious or even prohibitively time-consuming. Medical image segmentation is one such task which has been massively aided through the development of automated segmentation algorithms. Traditionally, human experts would have been required to manually annotate images to, for example, outline the location of a particular organ to enable further analysis of its shape or size. For many such tasks, automated algorithms have been shown to perform as well as humans while taking a fraction of the time. While the obsolescence of clinicians is unlikely, there will always be a role for computing in aiding and supporting the clinical workflow, allowing for better patient outcomes with fewer resources.

1.1 Image synthesis

One branch of machine learning is that of image synthesis. Image synthesis is the process of generating an image with a specific set of characteristics, usually learned by analysing an exemplar training dataset of images containing such characteristics. After this learning process, a synthetic image can be generated either from scratch or by changing the appearance of a source image.

Both of these applications are particularly useful in medical imaging. Medical images are

expensive and time-consuming to acquire. Being able to generate synthetic images with the desired set of characteristics from scratch is a cheap and powerful way of creating new data for multiple applications. Another feature almost unique to medical images is that the same region can be imaged in many different ways using different modalities or scanner settings. It is often the case that images from multiple modalities are required to get a complete understanding of a patient's condition. The ability to transform an image produced through one modality to one with the appearance of a different modality, therefore, has many applications, including roles in image registration, reconstruction and quality enhancement, and in clinical tasks such as segmentation and abnormality detection.

1.2 Contribution

In this thesis, we investigate several uses of image synthesis in medical imaging, with a focus on applications in neurodegenerative diseases. This work is divided into two background and five contribution chapters. Chapters 4, 5, and 6, 7, 8, are written so as to be able to be read largely independently by those familiar with the area, with occasional references to ideas discussed in earlier chapters. Detailed and relevant clinical and technical concepts are presented in Chapters 2 and 3 respectively, with relevant sections referred to throughout the thesis to provide further background detail. The main focus of these early chapters is to put the later work in context within both the health-care and computer vision domains and to provide a primer for those reading this thesis who are not familiar with the area. For the interested reader, we also share some early work towards extending Generative Adversarial Networks into 3D, which is discussed, along with other avenues for future work, in Chapter 9 and detailed in Appendix C. The following pages contain a brief summary of each chapter and their contributions.

Brain Lesion Segmentation through Image Synthesis and Outlier Detection

We first propose a novel image synthesis algorithm which can be used to transform images from one modality into another. We use this to identify pathology visible on Magnetic Resonance (MR) images by comparing real images to pathology-free synthetic images which have been produced from an image type in which the pathology is not visible. We find that the proposed method allows for unsupervised anomaly detection of white matter hyperintensities on MRI, with segmentation accuracies significantly higher than that achieved using three popular methods.

Published work

The work in this chapter has been published in the following articles:

Bowles, C., Qin, C., Ledig, C., Guerrero, R., Gunn, R., Hammers, A., Sakka, E., Dickie, D.A., Valds Hernndez, M., Royle, N., Wardlaw, J., Rhodius-Meester, H., Tijms, B., Lemstra, A.W., van der Flier, W., Barkhof, F., Scheltens, P., Rueckert, D., 2016, October. Pseudo-healthy image synthesis for white matter lesion segmentation. In International Workshop on Simulation and Synthesis in Medical Imaging (pp. 87-96). Springer, Cham.

Bowles, C., Qin, C., Guerrero, R., Gunn, R., Hammers, A., Dickie, D.A., Hernndez, M.V., Wardlaw, J. and Rueckert, D., 2017. Brain lesion segmentation through image synthesis and outlier detection. NeuroImage: Clinical, 16, pp.643-658.

Visual summary

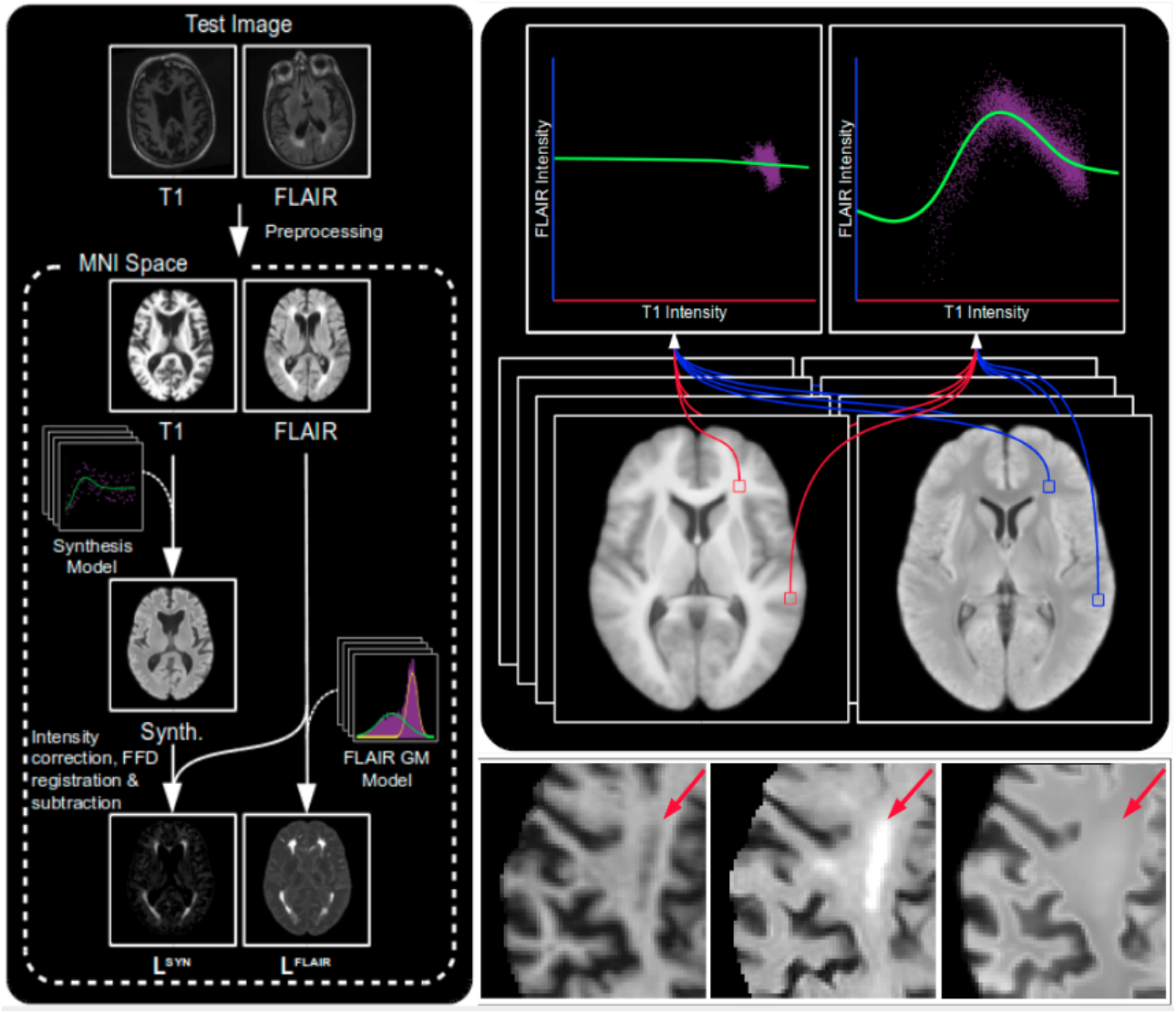


Figure 1.1: Summary of figures from Chapter 4 (Brain Lesion Segmentation through Image Synthesis and Outlier Detection). Left, overview of segmentation procedure. Top right, visualisation of proposed kernel regression based synthesis method. Bottom right, example of pseudo-healthy image synthesis. Please see Chapter 4 for full details.

GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks

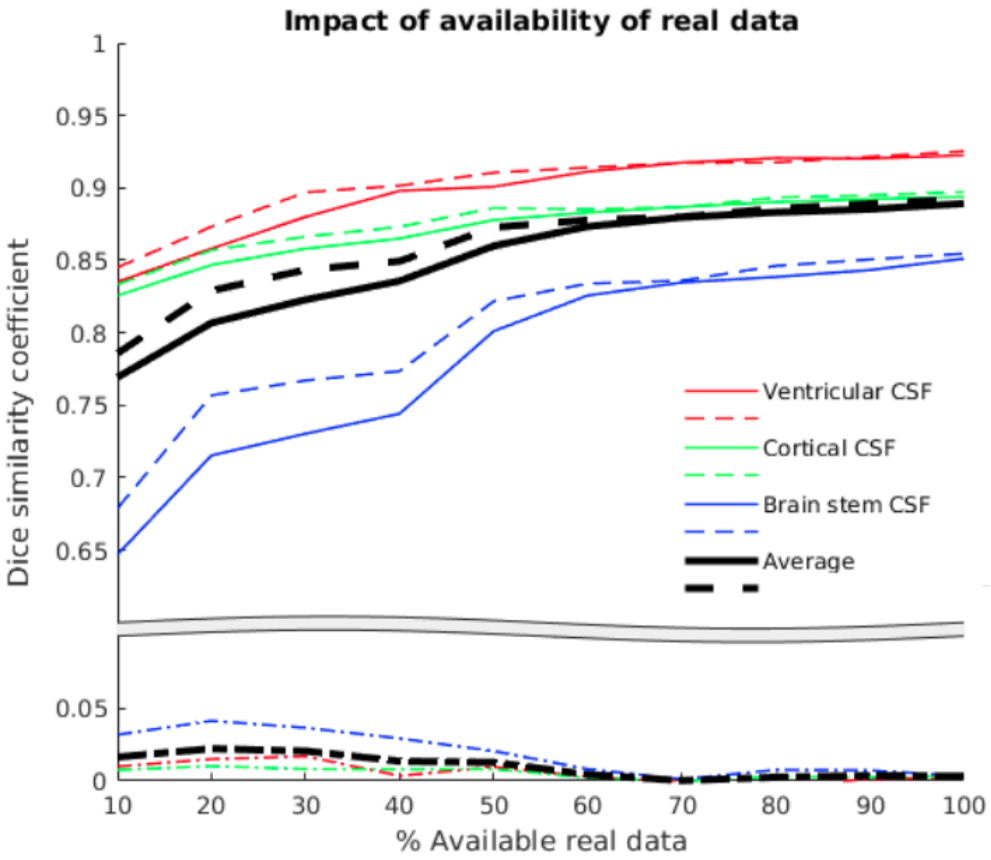
We next investigate whether synthetic labelled training data derived from Generative Adversarial Networks (GANs) can be used to expand existing manually labelled training datasets for segmentation algorithms. We evaluate the process across two datasets from different modalities (CT and MRI), using two segmentation algorithms, and under a variety of conditions where real training data is limited. We also compare performance with and without a traditional augmentation technique and perform a qualitative evaluation of the synthetic data. We find that across all experiments, GAN augmentation can lead to a significant improvement in segmentation accuracy where the amount of available data is severely limited, with the improvement diminishing as more real labelled data becomes available.

Published work

This work has been made available as a pre-print at:

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D.A., Hernndez, M.V., Wardlaw, J. and Rueckert, D., 2018. GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks. arXiv preprint arXiv:1810.10863.

Visual summary



	Available data		
	100%	50%	10%
No augmentation	88.1 (0.32)	85.0 (0.58)	75.1 (0.60)
GAN augmentation	88.4 (0.41)	85.6 (1.33)	76.3 (1.77)
Rotation augmentation	**88.9** (0.51)	**86.0** (0.50)	**76.9** (0.58)
GAN + Rotation augmentation	**89.3** (0.39)	**86.9** (0.36)	**78.4** (0.99)

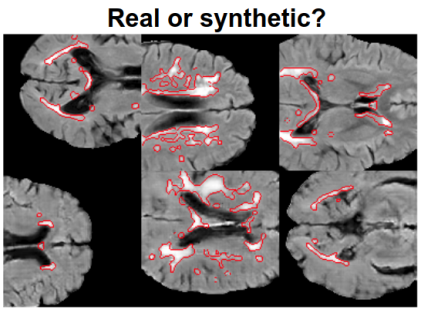


Figure 1.2: Summary of figures from Chapter 5 (GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks). Top, the effects of GAN augmentation where an increasing amount of real data are available. Bottom left, results of different augmentation approaches. Bottom right, a sample of mixed real and synthetic images. Please see Chapter 5 for full details.

GANsfer Learning: Combining labelled and unlabelled data for GAN based data augmentation

We build on the previous chapter by extending the synthetic data generation procedure to incorporate additional unlabelled data. We show that this enables the GAN to produce labelled images with greater anatomical variance. We use this to achieve both higher segmentation accuracies and improved Alzheimer's disease stratification scores from MR images by incorporating unlabelled elderly/pathological patient images into a dataset of predominantly young and healthy patient images.

Published work

The work in this Chapter was submitted for publication in IEEE Transactions on Medical Imaging and is in the process of revision. It has been made available as a pre-print at:

Bowles, C., Gunn, R., Hammers, A. and Rueckert, D., 2018. GANsfer Learning: Combining labelled and unlabelled data for GAN based data augmentation. arXiv preprint arXiv:1811.10669.

Visual summary

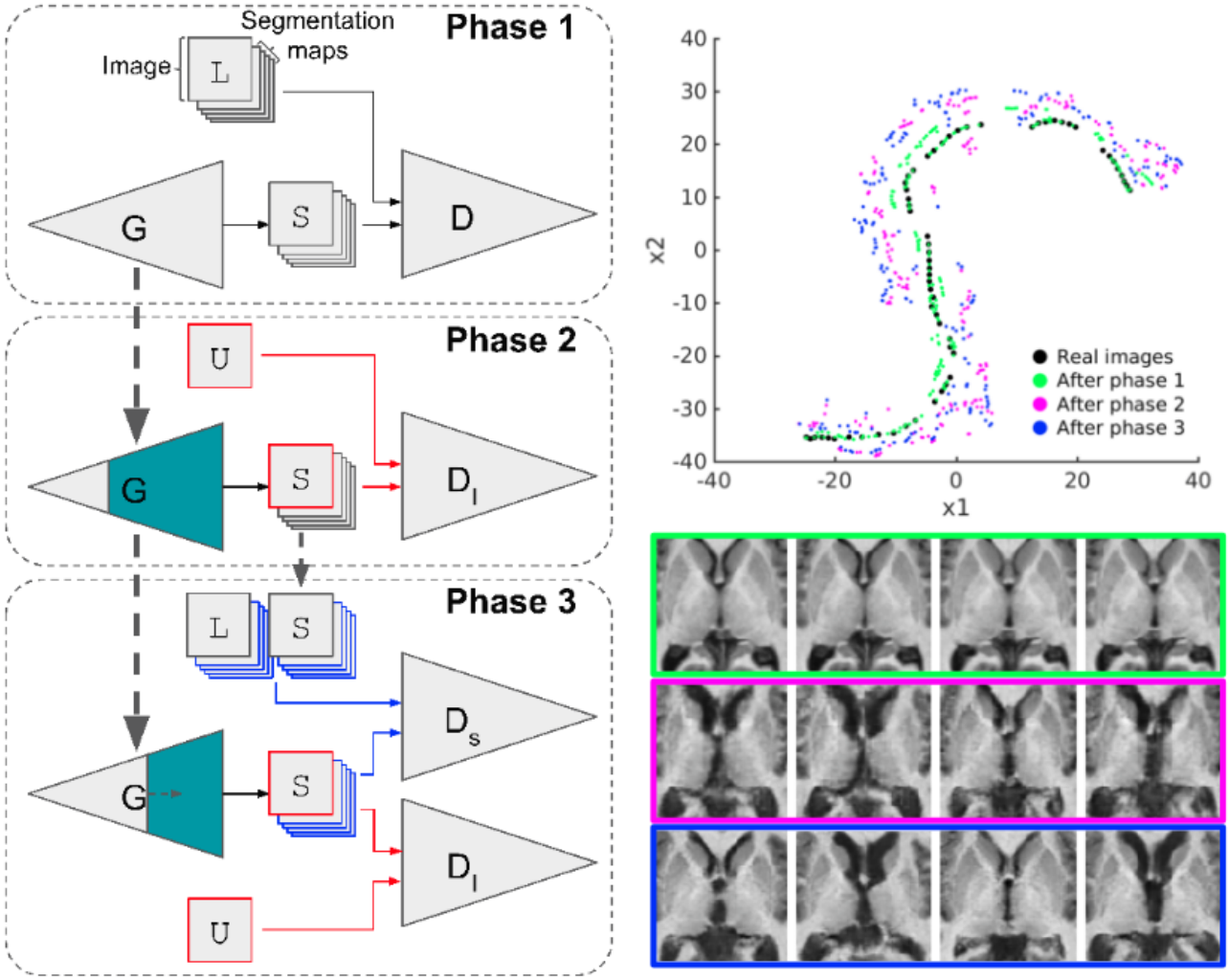


Figure 1.3: Summary of figures from Chapter 6 (GANsfer Learning: Combining labelled and unlabelled data for GAN based data augmentation). Left, the proposed three-stage method for combining labelled and unlabelled data in GAN training. Top right, a visualisation of the distribution of generated images after each phase of training. Bottom right, a sample of generated images after each phase. Please see Chapter 6 for full details.

Modelling the Progression of Alzheimer's Disease in MRI Using Generative Adversarial Networks - Parts A and B

Next, we develop a method which uses a class of deep neural networks called GANs to perform semantic editing of MR images. In Part A of this Chapter, we show how this allows for particular image characteristics, such as those associated with a disease, to be added or removed from real images. We demonstrate the method on a dataset of Alzheimer's Disease (AD) patients, showing how the common features of the disease can be added or removed from real images. In Part B we implement and extend the approach using a recently proposed high-resolution GAN formulation allowing for the procedure to be performed on entire 1mm isotropic slices, and demonstrate how the proposed method can also be used to discover associations between clinical variables and imaging data.

Published work

The work in part A of this Chapter has been published in the following article:

Bowles, C., Gunn, R., Hammers, A. and Rueckert, D., 2018, March. Modelling the progression of Alzheimer's disease in MRI using generative adversarial networks. In Medical Imaging 2018: Image Processing (Vol. 10574, p. 105741K). International Society for Optics and Photonics.

Visual summary

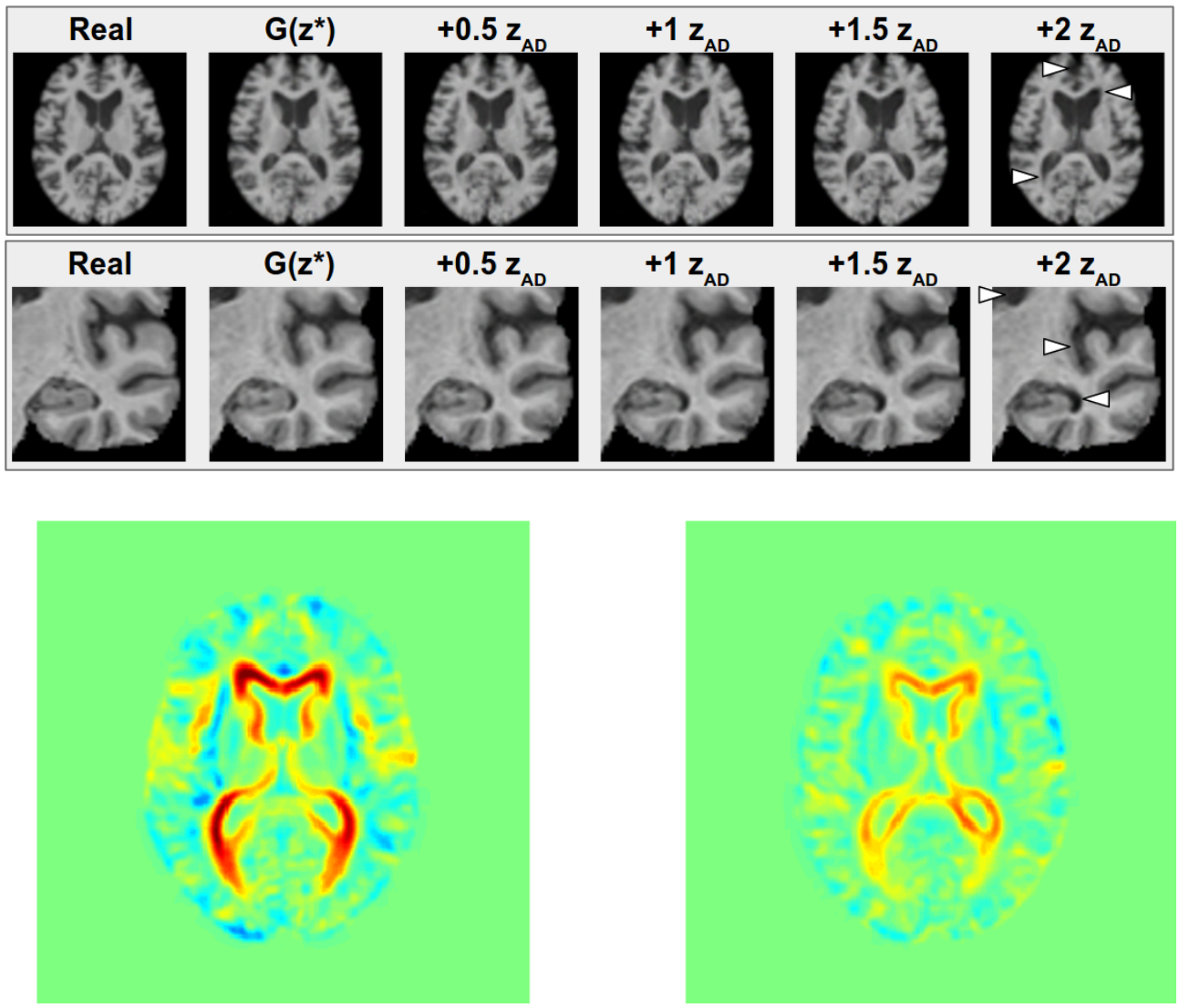
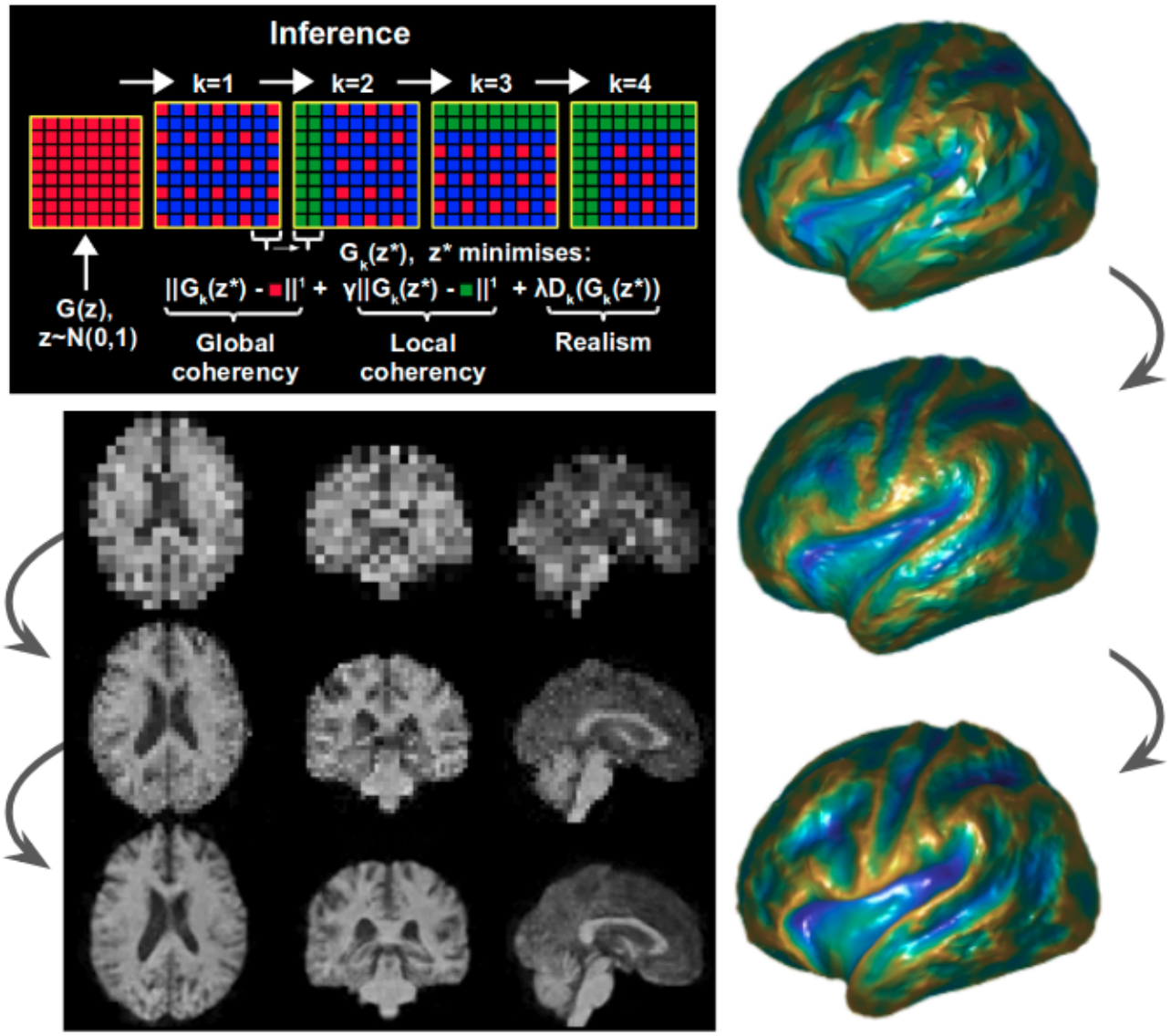


Figure 1.4: Summary of figures from Chapters 7 and 8 (Modelling the Progression of Alzheimer's Disease in MRI Using Generative Adversarial Networks - Parts A and B). Top, the effects of adding the characteristic features of AD to two regions of the brain through latent space image arithmetic. Bottom, the difference in predicted year-to-year atrophy levels in AD patients with and without the APOE4 allele, found through forming a model in latent space. Please see Chapters 7 and 8 for full details.

Future Work

Finally, we also share some ideas on how to effectively extend GANs into 3D, including some early results, as well as suggestions for other directions for future work in this fascinating field. We propose an approach to generating large or irregular image data by combining the output of multiple GANs in such a way as to enforce local and global coherency, and demonstrate this by creating 2mm isotropic MR volumes and high-resolution cortical surface data.



Chapter 2

Clinical Background

This chapter aims to place the content of this thesis within a clinical context. While the contributions of this thesis are primarily technical, it is important to keep in mind the clinical environment from which the proposed methods have been developed. It also provides a primer for readers who come from a more technical background, or perhaps from non-neuroimaging medical domains. This chapter starts with a basic overview of neuroanatomy, before progressing into the different ways this anatomy can be imaged. We then turn our attention to dementia, the main pathology explored in this thesis, from a purely clinical viewpoint: first describing the various pathological pathways, before looking at how they are usually diagnosed, the role of imaging within this process and the ethical considerations posed. We next look at small vessel disease and stroke, the major causes of vascular dementia, in more detail. We then review the Alzheimers Disease Neuroimaging Initiative (ADNI) and PredictND, two studies which have contributed to the growth of this field. Finally, we discuss some of the datasets which provide the data used in this thesis.

2.1 Neuroanatomy

The human brain is made up of two halves (hemispheres) and can be divided into three broad tissue types: White Matter (WM), Grey Matter (GM) and Cerebrospinal Fluid (CSF). GM can

be further divided into deep GM and cortical GM. As the name suggests, cortical GM makes up the cortex, the brain's outer layer with its characteristic wrinkled appearance containing ridges (gyri) and troughs (sulci). The deep GM comprises a number of structures which reside towards the centre of the brain.

GM structures are connected together by elongated nerve cells processes (axons). These connections form the WM and allow for communication between structures. Individual axons in the WM are surrounded by a tube of a fatty substance called myelin. This myelin sheath has the dual purpose of insulating the axon from those nearby and speeding up the transmission of signals along it.

Towards the centre of each hemisphere lie the lateral ventricles. These cavities contain and produce CSF, a fluid with several functions including the delivery and removal of nutrients and waste, suspension and protection of the brain from external forces, and insulation from pathogens. Beneath the lateral ventricles lie the third and fourth ventricles, which funnel the CSF towards the sub-arachnoid space, the area surrounding the cortex.

The brain is mostly symmetrical in structure, though not necessarily in function. While many pathologies affect both sides equally (bilateral), many others affect one half more than the other (unilateral). An overview of this anatomy can be seen in Figure 2.1.

2.2 Neuroimaging

Being able to acquire images from inside patients' brains is invaluable for both disease diagnosis and monitoring. Due to the multitude of structures and processes within the brain, there has emerged a host of different modalities for this acquisition. Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT), among others, all have an important role to play in understanding precisely how a patient's brain in functioning.

Such methods can be divided into two types - quantitative and qualitative. In quantitative

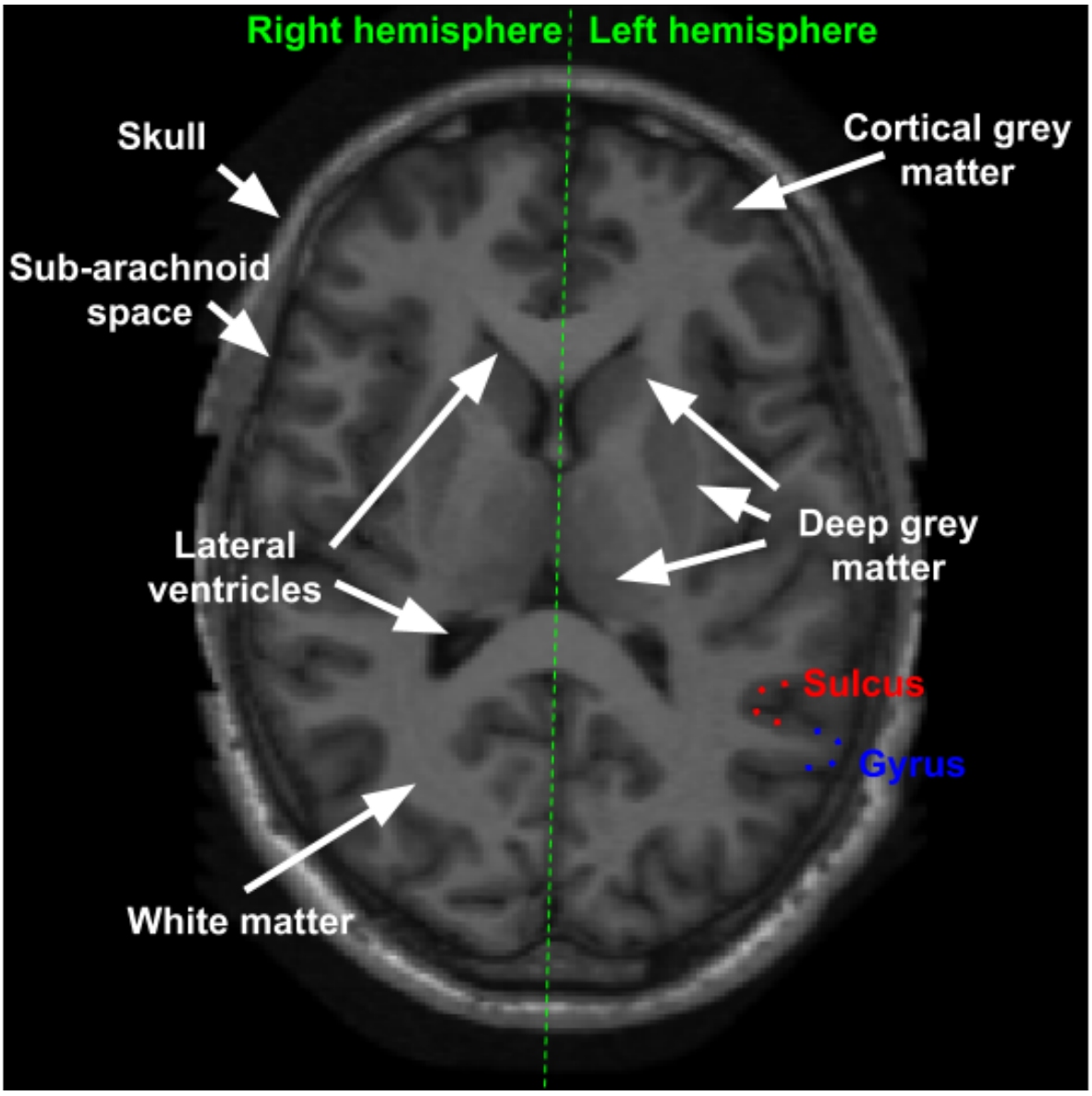


Figure 2.1: Overview of neuroanatomy as seen on a T_1-weighted magnetic resonance image.

imaging, the aim is to precisely measure the value of a property of the tissue being imaged. For example, CT directly measures the density of structures, while PET and SPECT aim to measure the concentration of radiation-emitting molecules in the tissue. In qualitative imaging, the images produced are not direct measures of a particular tissue property, but are, instead, measures derived from the interaction of a number of different properties. These approaches can be weighted towards a particular property, such is the case in a number of different MRI based methods, but they do not directly measure that property. While the majority of MRI based methods are qualitative, some, such as Diffusion Weighted Imaging (DWI) (described later), are quantitative. In addition, quantitative measures such as structure or blood volume can be extracted from qualitative images.

2.2.1 Magnetic Resonance Imaging

MRI is one of the most flexible tools in the modern-day clinician's repertoire. The technique allows for myriad different types of 3-dimensional (3D) images to be produced from a patient, providing information ranging from the exact locations of neural activity during different tasks, to the sizes, shapes and tissue make-up of a patient's brain structures. It has the advantage over PET, SPECT and CT of not involving any ionising radiation, and therefore exposes the patient to very little risk under regular usage.

At its heart, MRI measures a combination of two properties of the tissue it is imaging, T_1 and T_2. The exact combination of these two properties that is measured depends on the type of scan being performed, which is controlled by a set of Radio-Frequency (RF) pulses. The order and time between these pulses is known as an Magnetic Resonance (MR) sequence or protocol.

MRI protocols can be divided into two types, 2-dimensional (2D) and 3D. In 2D acquisitions, images are acquired as a series of 2D slices which are usually concatenated together to form a 3D volume. On the other hand, 3D acquisitions simultaneously acquire the entire 3D image volume. While 3D images can usually be used with few restrictions, care must be taken when working with images which have been acquired in 2D. Firstly, it is usual for 2D acquisitions do not have an isotropic resolution, and often low resolution in the through image plane.

Moreover, it is also common for such images to be acquired with gaps between the slices, with some regions of the image not measured. It is also important to consider that, in such cases, the value associated with a voxel is a weighted integral of the signal coming from that location. Any changes in signal strength occurring within the same voxel are therefore distributed over the entire voxel, which means small variations in the signal can be lost. The advantage of 2D acquisitions is that they are significantly faster, and are often necessary for protocols with long acquisition times. However, care must be taken when working with such images to be aware of their limitations.

The most used sequences are those which produce images which are more strongly weighted towards the T_1 property. These T_1-weighted images provide good contrast between tissue types and are mainly used to view a patient's anatomy.

T_2-weighted images, on the other hand, are particularly sensitive to water, which appears bright on images acquired under a T_2 sequence. This makes T_2 images good at identifying lesions, which often have high water content. However, the brain has a lot of naturally occurring water in CSF, the signal from which can be confused with that coming from lesions. Fluid-attenuated Inversion Recovery (FLAIR)([Hajnal et al., 1992]) images use a modified sequence which cancels out this confounding signal from CSF, providing much greater specificity for lesions. This, however, comes at a cost of resolution, meaning that, while T_1 images are often acquired at resolutions of $1mm^3$, FLAIR images are more commonly acquired in thick (eg. 5mm) slices.

Other sequences, such as those used for DWI and Functional Magnetic Resonance Imaging (fMRI) are sensitive to the movement of molecules. In the case of DWI this allows for local diffusion to be measured, while in fMRI, the destination of increased blood flow can be identified indicating areas of increased brain activity. Contrast MR is also used, where the patient is injected by or consumes a contrast agent with specific MR properties. The location of this contrast agent can then be imaged. One application of this is in the brain is Magnetic Resonance Angiography (MRA) where gadolinium is often used as a contrast agent which affects the T_1 properties of surrounding blood. This allows for detailed maps of the arterial structure in a

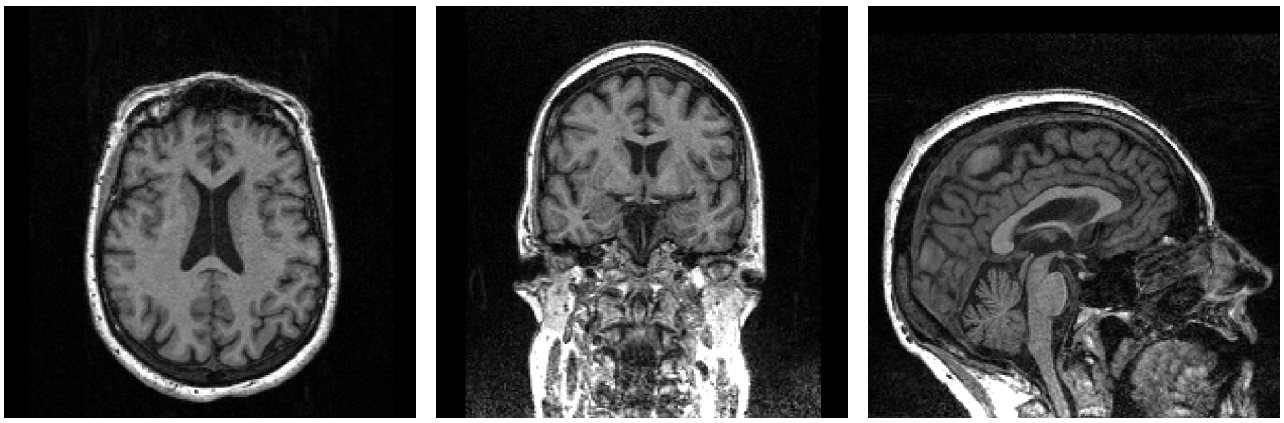


Figure 2.2: Example slices of a typical unprocessed T_1-weighted MR image taken from the ADNI dataset (see Section 2.5.1). Left to right: axial, coronal and sagittal views.

patient's brain to be produced. This is particularly useful in assessing stroke risk.

One of the drawbacks of MRI is its potential to produce artefacts in the acquired images. These are artificial effects that appear as a result of the failed acquisition or reconstruction of an MR image. These can be caused by a number of factors including: patient movement during scanning, the presence of metal objects such as fillings, the presence of air pockets such as in the sinuses, and attempting to reconstruct an image from too little information.

2.2.2 Computed Tomography

CT is another common method to acquire brain images. Like MR, it can produce 3D images covering the whole brain and can provide excellent contrast between different tissue types. CT is less flexible than MR, only providing one method for image acquisition. It is, however, cheaper and the equipment is more widely available.

CT acquisition involves transmitting X-rays through the patient onto a detector. Higher density tissues will absorb more of the X-rays, with fewer, therefore, reaching the detector. By acquiring these images from multiple directions around the Region of Interest (ROI), a detailed 3D image can be built describing the absolute densities of the tissue in that region.

CT is primarily a way to image anatomy and is particularly sensitive to regions with extreme densities such as bone and air. This makes it an effective way to image, for example, stroke

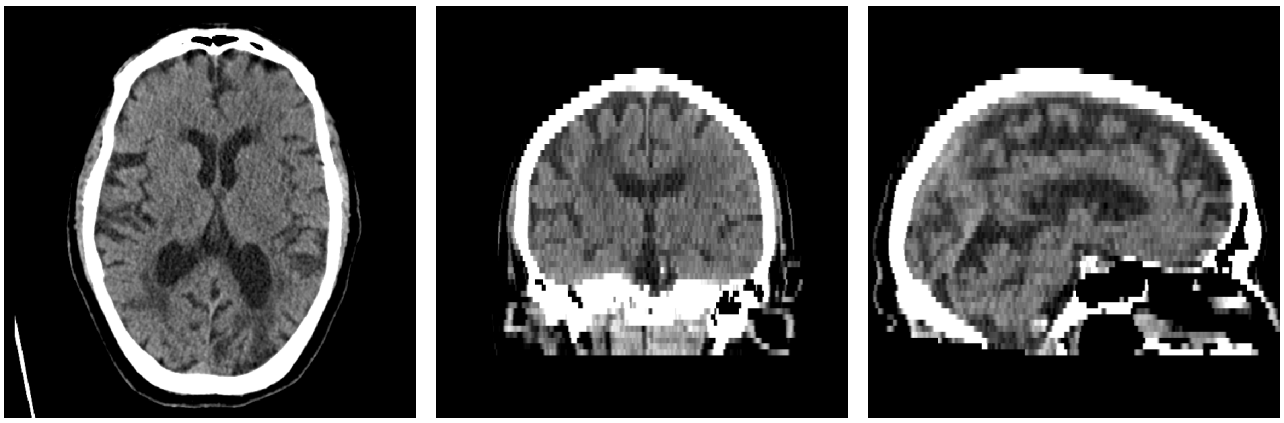


Figure 2.3: Example slices of a typical CT image. Intensities scaled to show brain tissue.

lesions which tend to have a different density to surrounding tissue. Contrast CT is also common. Similar to contrast MR, contrast CT involves using an extrinsic contrast agent to affect the properties of the tissue being imaged. CT angiography is therefore also an option for imaging blood vessels in the head, often using an iodine-based contrast agent.

2.2.3 Positron Emission Tomography

PET differs from the two methods already discussed in that it does not directly image anatomy. Instead, PET purely images the location of an injected radiotracer. This makes it useful for imaging the function of tissue. To acquire a PET image, the patient is first injected with a positron-emitting radioactive tracer which has been designed to be transported to a particular target of interest, eg. receptor protein. While there, the tracer will decay, releasing a positron which will travel a short distance before encountering an election and mutually annihilating. This process releases a pair of gamma waves in opposite directions which will pass through tissue and be detected using a ring of detectors outside the body. Exactly which detectors are activated defines a line along which the annihilation must have occurred. By detecting many such pairs, the areas in which the tracer has congregated can be localised.

One of the most common uses of PET is to locate cancer, and in particular, new areas to which cancer has spread. By using a tracer with similar properties to glucose (fluorodeoxyglucose (FDG)), an image can be acquired which indicates areas of increased metabolism - an indicator of cancer growth. In the brain, FDG accumulates in areas of neuronal activity (Figure 2.4),

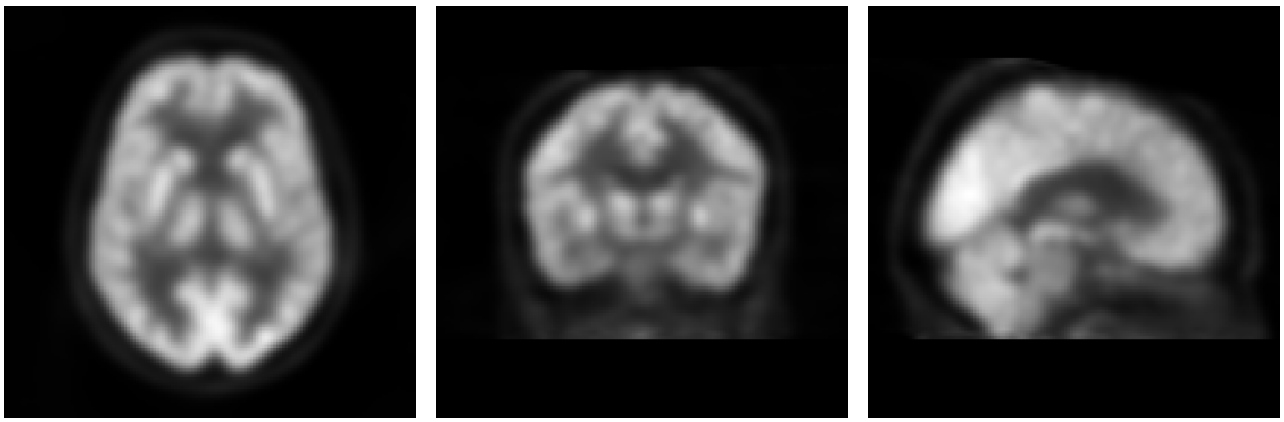


Figure 2.4: Example slices of a typical PET image.

allowing the reduced brain activity of dementia patients to be imaged.

2.2.4 Single Photon Emission Computed Tomography

SPECT is closely related to PET. It is also a functional imaging method which tracks the location of a tracer within the body. Unlike PET tracers which emit a positron, SPECT tracers decay directly into a gamma ray which is detected from outside the body. Detecting just a single gamma ray, as opposed to the co-incident rays in PET, means SPECT images have a lower spatial resolution than PET images. However, the radioisotopes used in SPECT tend to have a longer half-life and are therefore cheaper to produce and do not require specialist on-site equipment.

2.3 Dementia

Dementia is a term used to describe a number of diseases and conditions which result in neurological damage. These diseases lead to a progressive deterioration of cognitive functions until the sufferer can no longer function normally and may require round-the-clock care. In this section, we examine the global impact of dementia, describe some of the diseases which cause dementia and how they are diagnosed, and discuss some of the social and ethical questions which arise in dementia care and research. Unless otherwise referenced, the information in

this section comes from the Many Faces of Dementia course available online[1] led by Dr Tim Shakespeare, Alzheimers Research UK Fellow at the Dementia Research Centre, UCL Institute of Neurology.

2.3.1 Global Impact

Dementia primarily affects the elderly with the incidence rising exponentially with age. We live in a world with a rapidly ageing population, and as such, dementia is fast becoming one of the biggest and most expensive threats to global health. The estimated number of dementia sufferers worldwide is expected to rise from 46.8 million in 2015 to 74.7 million by 2030, with a corresponding increase in global spending from $818bn (1.09% GDP) to over $2tn [Prince, 2015]. There is currently no cure for dementia with most treatments only targeting the symptoms. The clear economic and human costs of this rapidly increasing incidence prompted the G8 Dementia Summit in 2013 to initiate a global effort to tackle dementia with the primary aim of identifying a cure or disease-modifying therapy by 2025 [Global Action Against Dementia, 2013].

2.3.2 Overview of types

There are four main diseases which are associated with dementia: Alzheimer's Disease (AD), Frontotemporal Dementia (FTD), Dementia with Lewy bodies (DLB) and Vascular Dementia (VD). The following is a summary of each.

Alzheimer's Disease

The most common cause of dementia is AD, accounting for 63% of dementia cases in the UK. The early symptoms of AD include personality changes, short term memory problems and a general struggle with everyday tasks such as handling money and navigation. As the disease progresses, these problems become worse. Patients develop problems recognising faces, become

[1] www.futurelearn.com/courses/faces-of-dementia/1

increasingly unable to carry out simple tasks such as getting dressed and short term memory problems get worse leading to the inability to learn new information. Depending on the parts of the brain most affected, patients may also experience hallucinations or delusions. Towards the end of the disease's progression, patients will be unable to function with any autonomy and require help with even the most basic tasks, eventually becoming unable to communicate entirely. Death is often caused by infection as opposed to the disease itself. For example, difficulty swallowing could lead to pneumonia, whilst incontinence could lead to urinary tract infections and urosepsis.

The pathological pathway of AD is characterised by the build-up of protein deposits in the brain. In particular, amyloid protein amyloid-β 1-42 ($A\beta_{42}$) begins to aggregate to form insoluble plaques, whilst incorrectly folded tau proteins create insoluble tangles. These protein build-ups eventually lead to cell death in the surrounding neurons.

Patients suffering from AD can be split into either sporadic or inherited cases. The latter, known as Familial Alzheimer's Disease (FAD), is rare, accounting for less than 1% of AD cases [Bateman et al., 2011], and is caused by inheriting a faulty copy of genes APP, PSEN1 or PSEN2. These genes are autosomal dominant, meaning that if a parent has one there is a 50% chance of it being passed on to an offspring who will then develop the disease. If inherited, the gene will trigger the pathological pathway described above and there is an almost certain chance that the child will develop AD later in life. Compared to patients with sporadic AD, those with FAD tend to develop symptoms earlier (aged 30-50).

Sporadic AD has no single discernible cause. A number of environmental factors such as diet and lifestyle, as well as genetics, play a role (see Table 2.1).

A variation of sporadic AD is Posterior Cortical Atrophy (PCA) where the back of the cortex, which is responsible for processing visual information, is primarily affected by the build-up of plaques and tangles. Patients with PCA will experience problems with their vision, even if their eyesight is perfect, and may, therefore, struggle with visual tasks such as driving or reading. Other early symptoms include difficulty with spelling and calculation, however, memory tends to be well preserved. As the disease progresses, these visual symptoms become worse until

Table 2.1: A summary of the different genes known to have an effect on a patient's risk of AD [Farrer, 1997, Guerreiro et al., 2013, Genin et al., 2011, Kamboh, 1995].

Genes	Frequency in population	Increased risk of AD
MS4a, CR1, PICALM, BIN1, CLU, CD2AP, CD33, EPHA1, ABCA 7	>50% have at least 1 variant	1-20%
1 copy of APOE4	25%	300%
TREM2	0.3%	300%
2 copies of APOE4	2%	800%

the patient becomes effectively blind. During the late stage of the disease, the patient may begin to lose other senses such as touch and eventually regress to the symptoms of late-stage AD. Patients with PCA usually develop symptoms at a younger age than in sporadic AD with current estimates of the prevalence of PCA suggesting that approximately 5% of patients who develop AD before the age of 65 have PCA.

Frontotemporal Dementia

FTD is the name given to a set of diseases which affect the frontal and temporal lobes of the brain. Accounting for 2% of dementia cases in the UK, FTD is relatively uncommon. As in AD, damage is caused to these areas through the build-up of proteins. However, in the case of FTD, these proteins are tau and ubiquitin. Memory is generally well preserved compared to AD, with symptoms related to behaviour and language being common.

The most common form of FTD is Behavioural Variant Frontotemporal Dementia (bvFTD). In bvFTD, the patient experiences changes in personality and behaviour. They may display compulsive actions and develop inappropriate social behaviour along with a loss of empathy. Changes in motivation and appetite are also common, especially overindulgence in sweet food. As with all forms of dementia, bvFTD is incurable. However, as most of the symptoms are behavioural, many treatment options do not involve medication, instead focusing on identifying

and avoiding behavioural triggers.

Semantic Dementia (SD), Progressive Non-Fluent Aphasia (PNFA) and Logopenic Aphasia (LPA) are three other forms of FTD. In all these conditions patients struggle with language, often being unable to find the right words to say. Patients with SD find problems with understanding the meaning of words and facts. PNFA primarily causes difficulty with speech, with patients often speaking slowly. Unlike in SD, the patient will still understand the meaning of the words but have difficulty in articulating them due to an inability to control the muscles in their mouth and face or to use the correct grammar. Similar to PNFA, patients with LPA will often pause during speech and struggle to repeat a sentence they have just heard. Whilst LPA is often classified under the header of FTD due to the parts of the brain which are affected, the pathological causes often more closely resemble AD, with the build-up of amyloid plaques and tau tangles. The type of FTD a patient may suffer from is heavily dependent on the part of the brain which is affected, with patients with language problems usually showing a greater level of atrophy on the left side of the brain than the right. In about 10% of FTD cases, the patient may also suffer from a Motor-neuron Disease (MND) such as Amyotrophic Lateral Sclerosis, or a movement disorder related to Parkinsons disease such as Corticobasal Syndrome or Progressive Supranuclear Palsy, and certain FTD co-morbidity combinations are more common than others. As with other dementias, death is usually caused by infection rather than the disease itself. Patients with co-morbid MND can die relatively soon, within 2-3 years, whereas some FTD patients may live 20 years or more.

FTD can be associated with a genetic cause, with three common autosomal dominant genes linked to the development of roughly a third of FTD cases. These genes are Microtubule-Associated Protein Tau, Progranulin, and C9orf72. Again, there is no direct relationship between gene and type of FTD developed, however certain genes are more likely to lead to particular combinations of FTD type and co-morbid disorders.

Dementia with Lewy Bodies

DLB accounts for 4% of dementia cases in the UK. Again, damage is caused by the aggregation of protein in the brain. In DLB, this protein is alpha-synuclein which aggregates into small structures, Lewy Bodies, inside brain cells. Patients with DLB often suffer from hallucinations and delusions, and it is strongly associated with Parkinson-like symptoms affecting movement. Whilst it's common for the symptoms of all dementia types to fluctuate over a few days, in DLB these fluctuations can be much more rapid, sometimes over a duration of minutes.

The relationship between DLB and Parkinson's disease is very close. The order in which symptoms develop dictates whether a patient is given a diagnosis of DLB or a separate diagnosis of Parkinson's Disease with Dementia (PDD).

Vascular Dementia

Accounting for 17% of dementia cases in the UK, VD is the second most common cause of dementia after AD. Unlike the dementias described so far, VD is not associated with abnormal protein build-up in the brain. Instead, damage is caused to brain cells by a reduction in blood supply as a result of damaged blood vessels [Kurz, 2001]. There are two major causes of damage. The first is stroke, where the blood supply to part of the brain is interrupted, either as a result of a blockage (ischaemic stroke) or a bleed (haemorrhagic stroke), both of which can cause permanent damage. The second cause is Small Vessel Disease (SVD) in which the walls of the blood vessels in the brain become harder and thicker, thereby restricting blood supply. The symptoms of VD are largely dependent on the size, location and number of strokes, or degree of SVD. Risk factors for both stroke and SVD include high cholesterol, high blood pressure, diabetes and smoking.

Mixed Dementia

Up to 10% of dementia cases in the UK can be attributed to a mixture of two separate pathologies. The most common of these combinations being AD and VD. However, due to completely

unrelated pathological pathways, VD can be co-morbid with any other dementia type.

2.3.3 Clinical diagnosis of dementia

It is important for researchers, particularly those without a clinical background, to understand the clinical protocols used when diagnosing patients with dementia. This allows for genuine needs to be identified and solutions proposed.

The biggest risk factor for almost all types of dementia is age, with the vast majority of cases occurring in the elderly. All cases of dementia are diagnosed based on patient history and a key symptom is a change in cognitive function. However, this can occur slowly, often imperceptibly, and with the patient unaware of any changes. Referral to a specialised dementia clinic is often required as it can be difficult to identify these subtle changes during short GP appointments. It is vitally important for the diagnosing clinician to speak to both the patient and their family to gain an impression of the patient's changes over time. As well as patient history, a number of biomarkers exist for each type of dementia. Biomarkers are measurements of a pathological process which can be extracted using techniques such as imaging or biopsy. These measurements can be compared to predefined benchmarks to provide evidence for or against a particular diagnosis.

Alzheimer's Disease

According to the clinical guidelines current during the work on this thesis [McKhann et al., 2011], a diagnosis of AD is initially made through the patient's history. Information such as the time and duration of symptom onset and the types of cognitive deficits exhibited will be recorded and used to form a diagnosis. Providing there is no evidence for any other type of dementia, a diagnosis of probable AD may then be provided. An increased level of certainty in the diagnosis can be provided through a documented record of cognitive decline through formal examinations or by genetic testing (in the case of FAD). Genetic testing is usually only carried out if there is a strong history of early-onset AD in the patient's family. Further levels of confidence in the

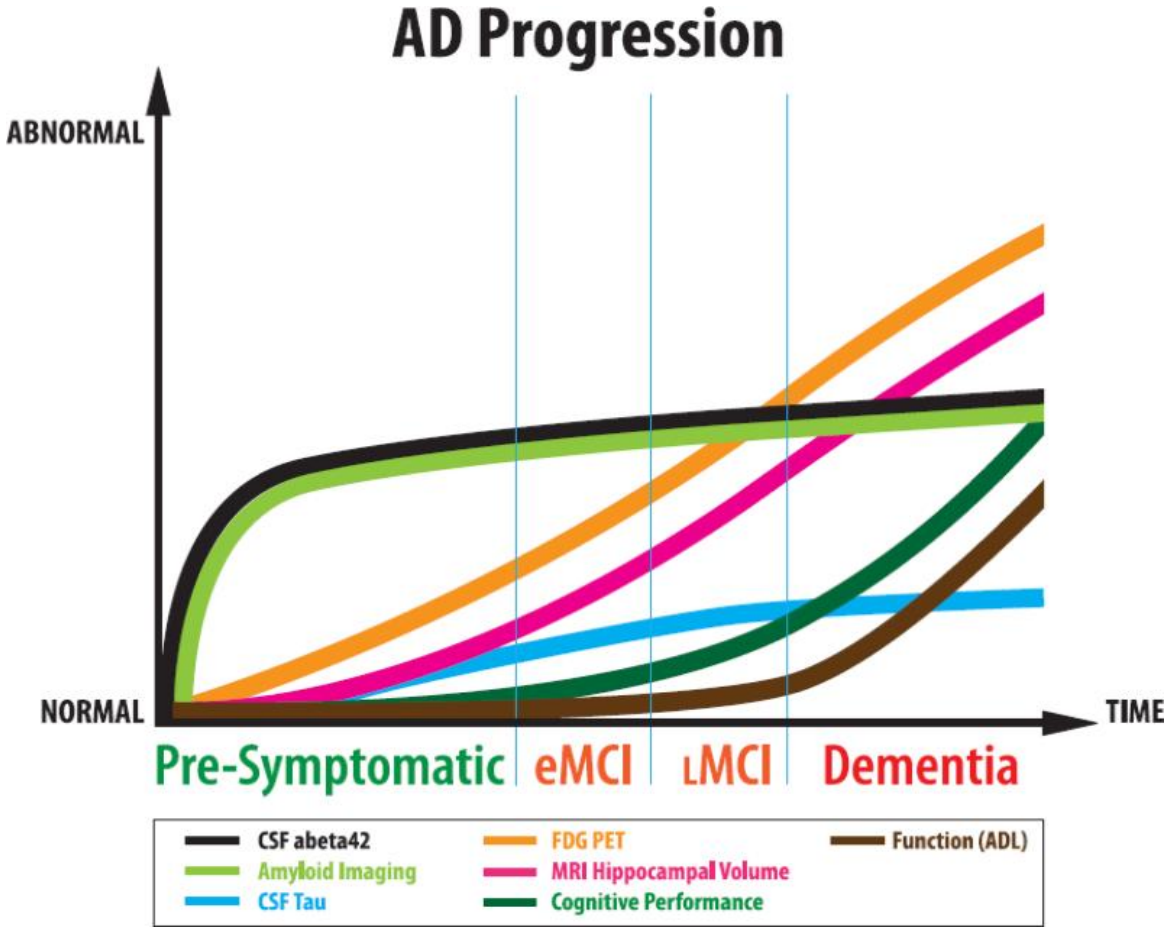


Figure 2.5: Theorised changes in key biomarkers through the progression of AD. Figure from [Aisen et al., 2010], adapted from [Jack et al., 2010], used with permission.

diagnosis can then be added through the collection of various biomarkers. These can include the collection of CSF to examine the levels of Aβ_{42} and Tau proteins, the acquisition of structural MRI images to determine a level of tissue atrophy and the acquisition of PET images to examine the levels of amyloid build-up or glucose metabolism. Figure 2.5 shows how a number of biomarkers change from normal to abnormal during the progression of AD. The earlier the biomarker changes from normal to abnormal, the earlier it can be used to identify a potential case of AD. However, since cognitive function is the last to change, the patient is often not referred to the clinic until the later stages, even when tests exist to potentially identify the developing disease earlier.

Frontotemporal Dementia

As a result of the many diseases which fall under the umbrella of FTD, diagnosis can be difficult. As with AD, the first step is to establish a full patient history both from the patient and their family or friends. Gaining information from the people close to the patient is especially important in bvFTD as the patient may believe that they are behaving in a normal way, even if they are not. Appendix A.1 shows a list of behavioural features of bvFTD, along with some examples. Diagnosing clinicians will look for these signs to identify the patient as having bvFTD, as opposed to other variants of FTD or other dementias. For the differential diagnosis of the other FTD variants, a flow chart may be used, such as the one seen in Appendix A.1. A key step in this chart which should be noted is that brain imaging is carried out, which ensures patients suffering from non-degenerative pathologies, such as a tumour, are correctly identified and not misdiagnosed [Warren et al., 2013].

Dementia with Lewy Bodies

Due to the similarity of symptoms, differentiating between DLB and PDD is difficult. In practice, the diagnosis of PDD is given only in cases where the Parkinson's symptoms, primarily motor symptoms, appeared over a year before the onset of cognitive changes. Otherwise, a diagnosis of DLB is given.

Vascular Dementia

One of the most commonly used diagnostic guidelines for VD is the NINDS-AIREN criteria [Román et al., 1993]. These very specific guidelines aim to establish that the patient has a sufficient level of cognitive impairment, evidence of cerebrovascular disease and that there exists a temporal relationship between the two. In other words, cognitive decline began after a cerebrovascular event or that there has been an abrupt or stepwise decline.

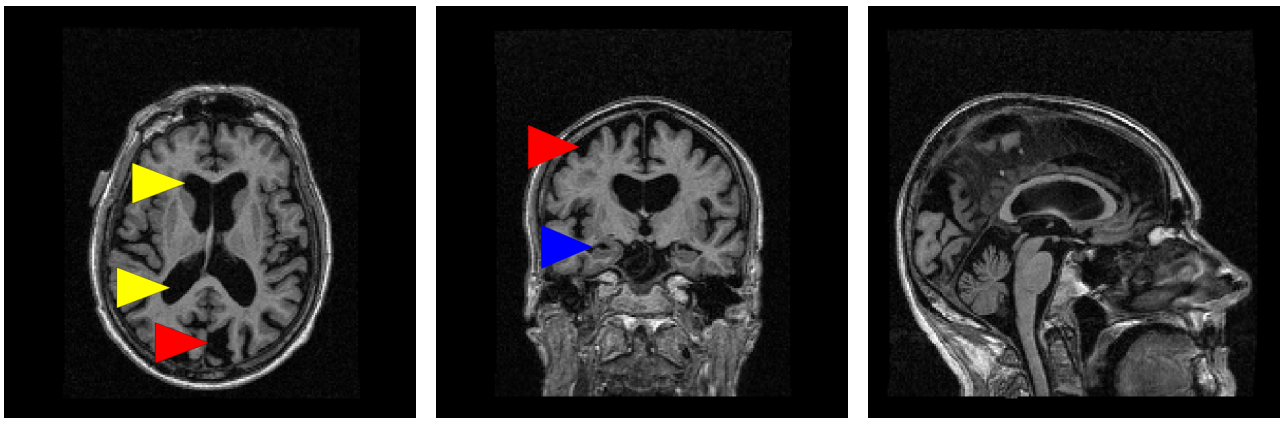


Figure 2.6: Example slices from a patient showing the typical features of AD. Enlarged ventricles (yellow), cortical atrophy (red) and hippocampal atrophy (blue).

Mixed Dementia

Mixed dementia cases can be challenging to diagnose, as the patient will often exhibit symptoms from both diseases. Usually, if an MRI scan fulfils the imaging components of the NINDS-AIREN criteria for probable VD and the patient also shows sufficient evidence for another type of dementia, then a diagnosis of mixed dementia will be given. If there is some evidence of VD, but not enough to satisfy the NINDS-AIREN criteria, then a diagnosis could be, in the case of AD co-morbidity, "Alzheimer's disease with a vascular component", or similar.

2.3.4 Role of imaging

As discussed previously, a common effect of dementia is the shrinking of the patient's brain. The exact location and degree of this varies between the conditions and can be an important distinguishing factor in their differential diagnosis. Whilst almost universally present, this shrinkage can be subtle, especially in the early stages of the disease when forming a diagnosis is the most challenging. Highly detailed images are required to identify these small changes which are readily provided by T_1-weighted MR images (see Figure 2.6).

Vascular dementia is less commonly associated with significant tissue loss and as such is unlikely to be identified by measuring structural volumes. However, SVD and stroke, the common precursors to VD, can be identified and quantified using MRI. Whilst sometimes visible

on T_1-weighted MR images, these lesions are much more prominent on FLAIR images. The appearance and classification of these lesions are discussed further in Section 2.4.

2.3.5 Ethical considerations in dementia diagnosis

Genetic testing, as often used to confirm cases of FAD and other inherited dementias, always poses challenging ethical questions. In the case of autosomal dominantly inherited genes, it is important to consider whether the patient wants to undergo genetic screening, as being a carrier almost guarantees the development of dementia in later life. Not only does the discovery of one of these genes impact the patient, but it also impacts their family as 50% of the patient's siblings, and one of their parents, will also be highly likely to develop dementia. Genetic counselling is often made available in cases such as these to give the opportunity for the patient and their family to make an informed decision before undergoing any genetic testing.

Whilst there are a number of different diseases which cause dementia, many symptoms are shared across each of them. These include memory loss, planning and organisational difficulties and personality changes. As a result, many sufferers require a great deal of care and support, often from family members, the degree and scope of which will vary between cases depending on the underlying disease and its progression. This can be particularly difficult for the carer as the patient can often become unrecognisable from the person they were before.

Not only is there a social imperative that carers are looked after as well as the patient, there are also economic benefits to considering how carers are impacted by clinical decisions made for the patient. Of the $818bn global cost of dementia, $327bn (40%) is attributed to the informal care of dementia patients by family carers [Prince, 2015]. This includes not only an estimate of lost income due to lost time working but also the cost of medical conditions resulting from the additional stress and anxiety being a carer can cause.

From a research point of view, it is important to consider what impact different fields of research may have on the carer as well as the patient. This leads to an interesting debate. Should time and money be spent on an accurate diagnosis of dementia when it does little to change the

prognosis of patients, and the resources could instead be spent improving the lives of people with dementia in other ways? This is an important question to consider, though its discussion is beyond the scope of this work.

2.4 Cerebral Small Vessel Disease and Stroke

Cerebral small vessel disease is common in the elderly with severe cases leading to cognitive impairment in the form of VD. While the cause of SVD is not always clear, risk factors include age, smoking, and elevated blood pressure [van Dijk et al., 2008]. SVD can manifest in a number of ways [Wardlaw et al., 2013], usually as a result of intrinsic brain small vessel abnormalities leading to an inadequate blood supply (ischaemia). Brain tissue damaged as a result of ischaemia presents as bright (hyperintense) on T_2-weighted MR images (eg FLAIR) and often dark (hypointense) on T_1-weighted images, see Figure 2.7. This is because ischaemia, and associated demyelination, increases local water content in the brain. This causes a lower T_1 signal, and higher T_2 signal. This is SVD can also lead to lacunes (fluid filled cavities <20mm diameter with an MR appearance similar to CSF, sometimes with a T_2 hyperintense ring); enlarged perivascular spaces (extra-cerebral fluid around vessels, $<$ 2 mm diameter, similar MR appearance to small lacunes without T_2 hyperintense ring); and cerebral microbleeds (leakage of blood cells into perivascular tissue, visible as <10mm diameter hypointensity on T_2*-weighted and susceptibility weighted MR sequences) [Wardlaw et al., 2013].

Most attempts to automatically quantify SVD [Caligiuri et al., 2015] have focused on the accurate segmentation of hyperintense lesions within the WM on FLAIR images. FLAIR is the most useful MR sequence for the detection of these lesions as it is a T_2-weighted sequence in which signals from confounding sources of hyperintensity, primarily CSF, are cancelled out. There has been comparatively little work on identifying the other manifestations of SVD such as lacunes [Ghafoorian et al., 2017], perivascular spaces [Del C. Valdes Hernandez et al., 2013] and microbleeds [Kuijf et al., 2012].

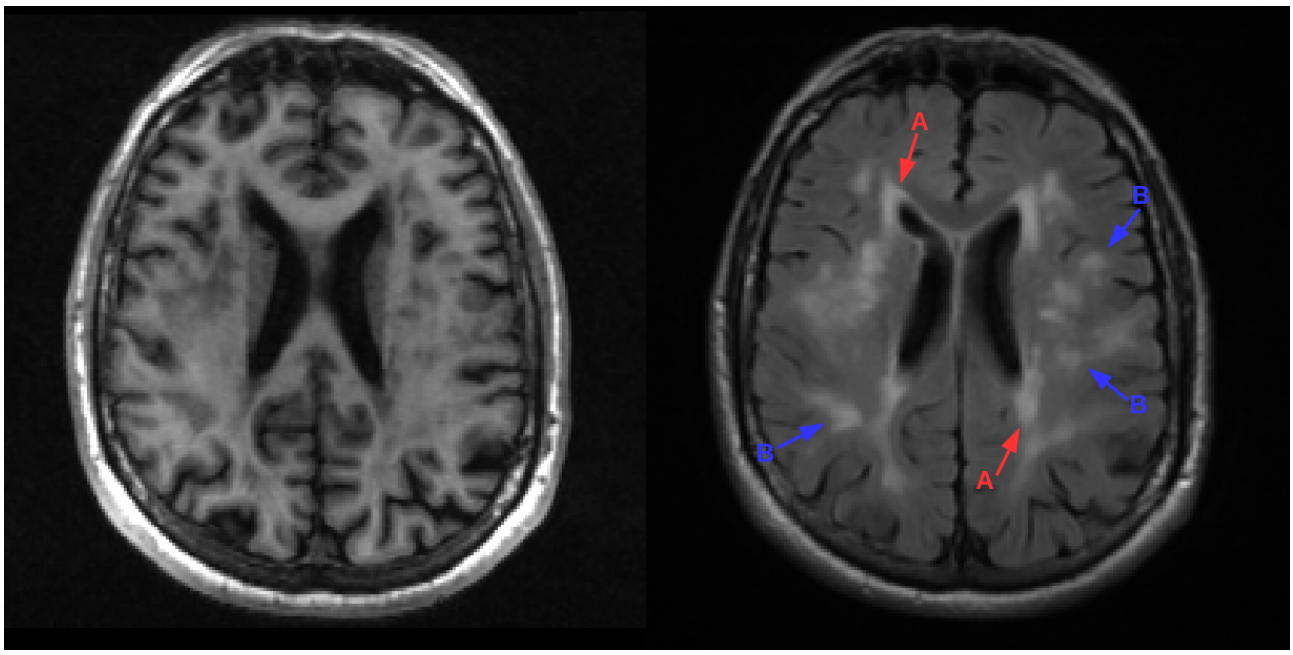


Figure 2.7: T_1 (left) and FLAIR (right) image of a subject with periventricular (A) and deep (B) white matter lesions. Note that pathology is more visible on the FLAIR image than it is on the T_1 image.

2.4.1 A note on terminology

The terminology and definitions surrounding SVD and associated imaging features can vary significantly between studies [Wardlaw et al., 2013]. To avoid confusion, this section defines the following relevant terms explicitly in line with those given by Wardlaw et al. with examples of each shown in Figure 2.8. The term White Matter Hyperintensity of Presumed Vascular Origin (WMH_{pvo}) refers to the lesions within the WM which appear hyperintense on T_2-weighted MRI (including FLAIR) which are often present in images of older people. WMH_{pvo} are often symmetrical and their cause is unclear. The term Recent Small Subcortical Infarct (RSSI) refers to a T_2 / DWI hyperintense region indicating a region of recent tissue death (infarction). An RSSI will evolve into either a lacunar cavity (T_1 / T_2 hypointense "space", usually with a T_2 hyperintense ring) or T_2 hyperintensity. The term WMH is used to include all T_2 hyperintensities caused by WMH_{pvo}, RSSIs, RSSIs which have evolved into T_2 hyperintensity and the T_2 hyperintense areas around lacunar cavities. Finally, the term cortical infarct is used to refer to T_2 hyperintense regions which appear wholly or partly in the cortical GM.

Whilst Multiple Sclerosis (MS) lesions also manifest as hyperintense regions within the WM

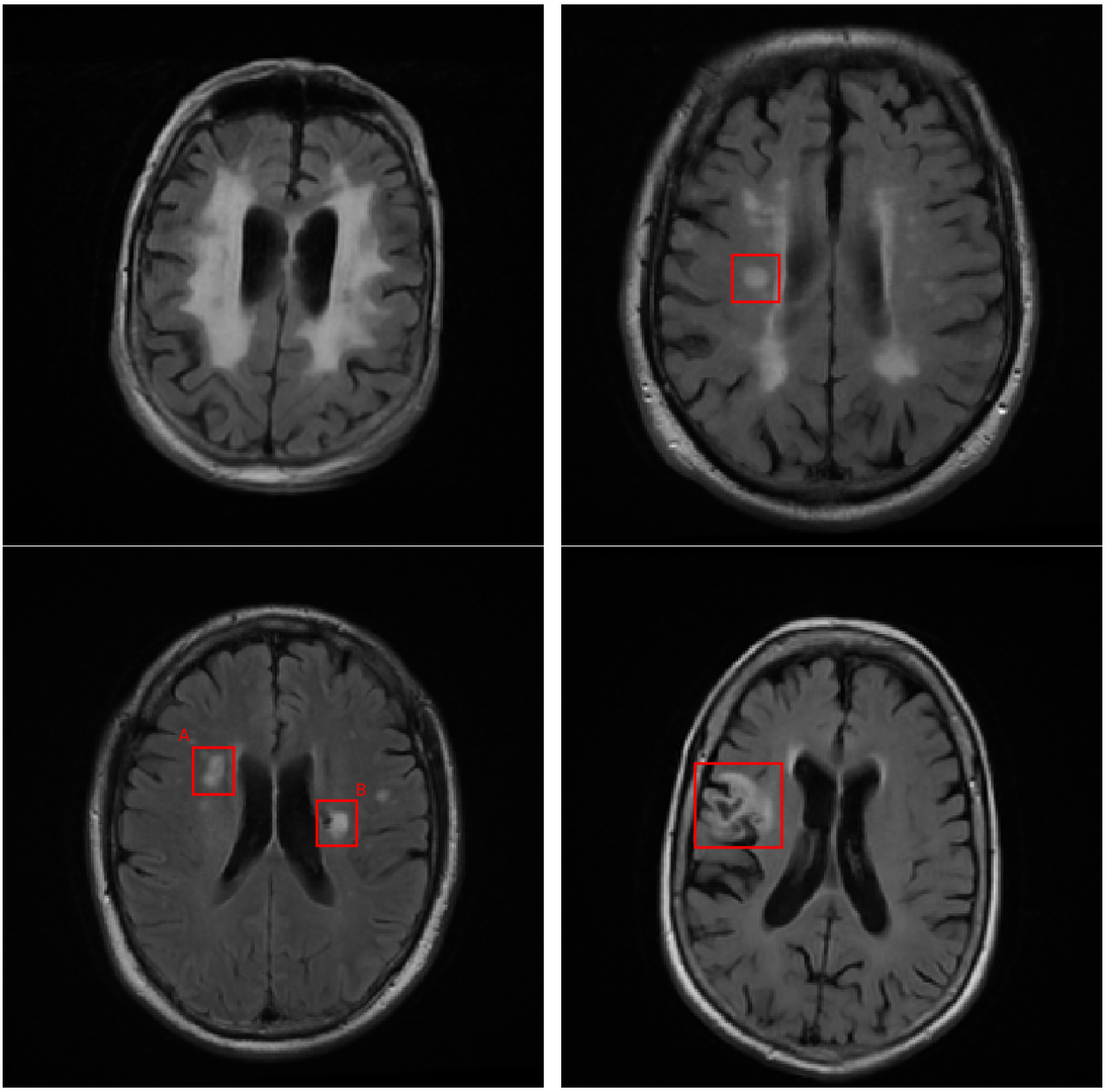


Figure 2.8: Examples of different hyperintensities relating to SVD. Top left: White matter hyperintensity of presumed vascular origin. Top right: Recent small subcortical infarct. Bottom left: A: Evolution of a recent small subcortical infarct into a T_2 hyperintensity, B: Lacunar cavity forming at the edge of a White Matter Hyperintensity (WMH) of unclear origin. Bottom right: Cortical infarct.

on FLAIR [Polman et al., 2011], work in this thesis is primarily focused on the vascular causes mentioned above, hence the definition of WMH is reserved to these, and MS induced hyperintensities are referred to separately as MS lesions where necessary. No other cause of T_2 hyperintensity (eg cancer, traumatic brain injury) is discussed in this thesis, or present in any experiments.

2.5 Dementia Neuroimaging Studies

2.5.1 The Alzheimers Disease Neuroimaging Initiative (ADNI)

ADNI [Mueller et al., 2005] is a large scale AD study launched in 2003 following the progression of patients over several years (longitudinal). Its aim has been to investigate whether the diagnosis and onset prediction of AD can be improved via the use of imaging data, particularly MRI and PET, combined with other clinical and cognitive biomarkers. Due to its success, this remit was extended through two further studies, Alzheimers Disease Neuroimaging Initiative - Grand Opportunity (ADNI-GO) [Alzheimer's Disease Neuroimaging Initiative, 2010] and Alzheimers Disease Neuroimaging Initiative - 2 (ADNI-2) [Weiner, 2014]. The former expanded the study to include more patients from across the spectrum of AD, from early cognitive impairment, through to confirmed AD. This, along with additional standardised tests, allowed for greater insight into the earlier stages of the disease and symptom development. ADNI-2 secured funding to continue following the patients already enrolled in the first two studies, as well as to recruit more patients from across the whole spectrum of the disease. Alzheimers Disease Neuroimaging Initiative - 3 (ADNI-3) [Weiner et al., 2017] began in August 2016 and is currently ongoing. It continues to follow patients enrolled in ADNI-2 and aims to recruit up to 1200 new subjects. It will introduce additional PET data in the form of both tau and amyloid imaging, as well as a stronger focus on using MR for connectivity analysis.

The ADNI database currently holds detailed longitudinal data on over 800 patients, including: patient/family history, imaging data (MR and PET), neuropsychological assessment scores,

genetic information and CSF measurements (see `http://adni.loni.usc.edu/data-samples/adni-data-inventory` for a full data description).

The ongoing success of ADNI is evident as it enters its 15^{th} year. It has been credited with many important findings, with the rubrics of the later studies heavily influenced by the discoveries of those before them. The opening up of such a large amount of standardised data to the community has undoubtedly led to a higher rate of discoveries than would otherwise have been possible. This is perhaps made most evident when one considers a comprehensive 2013 review of studies based upon the ADNI data contained over 100 pages and 300 references [Weiner et al., 2013].

Combining heterogeneous data from multiple studies in an unbiased way can be a considerable source of work and is often impossible. Therefore, much of the attraction of the ADNI dataset to researchers is the high degree of standardisation allowing for large amounts of data to be analysed together with confidence and significant conclusions to be made. Many of these conclusions have come in the search for new biomarkers [Hampel et al., 2008, Shaw et al., 2007, Clark et al., 2007]. One of the biggest accomplishments of ADNI has been in the discovery of CSF-based protein biomarkers for AD - Aβ_{42} and total tau [Shaw et al., 2009].

These biomarkers, among others, have been incorporated into machine learning systems for diagnostic purposes. Such large scale homogeneous datasets are perfect for machine learning classification approaches, as evidenced by the large number of such studies, for example: [Misra et al., 2009, Gray et al., 2013, Casanova et al., 2011, Abdulkadir et al., 2011]. One study, called PredictAD [Mattila et al., 2011], demonstrated how such a classifier was able to improve AD prediction [Liu et al., 2013] by being incorporated into a comprehensive decision support system.

Use of Imaging Data

The increased prevalence of PET and MR scanners has accelerated the interest in image-derived biomarkers. Alongside clinical and neurochemical biomarkers, imaging biomarkers are increasingly being used in the diagnosis and stratification of dementia. One such biomarker is

hippocampal volume, first tested using volumetry in 1999 [Jack et al., 1999], its potential was later confirmed using data from the ADNI cohorts [Schuff et al., 2009].

One of the challenges faced when identifying imaging biomarkers is inhomogeneity within and between datasets. Different scanners and acquisition protocols can lead to images with substantially different appearances. This is a challenge for many automated algorithms which often rely on intensity differences between structures as features for segmentation or classification. Comparisons between methods can, therefore, be a challenge - the best algorithm on one dataset may not be the best on another. Many segmentation methods have been proposed for MR images [Balafar et al., 2010], with many also applied to investigate dementia. However, the choice of algorithm varies between studies, leading to further barriers to direct inter-study comparisons. Multi-atlas segmentation has been shown to perform well across different acquisition protocols [Babalola et al., 2008], and is, therefore, a valuable tool in the search for imaging biomarkers. A step towards this has already been made in the publication of a repository of ADNI images segmented using the MAPER algorithm [Heckemann et al., 2010]. This has also very recently been added to in [Ledig et al., 2018a], along with morphological analysis [Ledig et al., 2018b] using an updated version of MAPER, called Multi-Atlas-Label Propagation with Expectation-Maximisation based refinement (MALPEM) [Ledig et al., 2015].

Differential Diagnosis

ADNI has lead to great strides forward in the field of AD diagnosis, stratification, and monitoring. However, the differential diagnosis of dementia has remained comparatively unexplored. [Shaw et al., 2007] highlighted the need for biomarkers which can separate cases of AD from other dementias. While the pattern of glucose metabolism shown using FDG-PET has been proposed for this purpose [Kannan et al., 2009, Herholz, 1995, Mazziotta et al., 1992], its further study has been limited due to comparatively small sample sizes [Foster et al., 2007, Dukart et al., 2011, Hoffman et al., 2000]. Larger studies into the use of FDG-PET have instead focused on AD diagnosis [Silverman et al., 2001]. In addition to FDG-PET, atrophy levels of cerebral structures have also been investigated as being potentially useful in discrimi-

nating between the different dementia-causing diseases [Vemuri et al., 2011, Galton et al., 2001, Burton et al., 2002].

2.5.2 PredictND

Large standardised datasets such as ADNI can provide a lot of data upon which machine learning models can be trained, allowing for hypotheses to be tested without the relevant information becoming obfuscated by noise. However, one must consider the implications of this when applying models trained on such datasets to the more heterogeneous data usually found in clinical practice. A model trained on standardised data will not generalise as well to other datasets as one trained on data from a variety of sources, as it is likely to overfit its training set. This can lead to over-confidence in the performance of a particular model on one dataset where it was trained on another. To build confidence in a model, it is therefore important to evaluate and, if necessary, train on data from more than one source. As such, studies which combine data from multiple sites with different clinical protocols are valuable to both develop algorithms and assess their generalisation capabilities.

PredictND is one such study. Launched in 2014, PredictND is a 4.2 million European project with the aim of investigating all causes of dementia.

It has four objectives:

Scientific objectives:

- To develop an IT-supported clinical protocol for enabling early and objective differential diagnostics of neurodegenerative diseases based on the principles of data-driven evidence-based medicine.
- To develop a low-cost battery of tests for early detection of cognitive change.

Technical objectives:

- To develop a decision support software tool to be used in clinical workflows for differential diagnostics of neurodegenerative diseases.
- To develop an ICT ecosystem and objective diagnostics of neurodegenerative diseases.

The project involved the recruitment of up to 800 patients from across four clinical sites: Kuopio University Hospital (Finland), Region Hovestaden (Copenhagen, Denmark), VU Medical Center (Amsterdam, the Netherlands) and Hospital Perugia (Italy), from across the spectrum of dementia-causing diseases. Such a large dataset offers a highly useful platform upon which methods can be developed and tested. The multicenter nature of the study means that, while acquisition protocols remain similar between sites, there were multiple scanners and staff members involved in acquiring these datasets. As a result, the final collection of data is heterogeneous and therefore provides the opportunity for the ability of methods to generalise to be explored. An early study [Koikkalainen et al., 2016] investigated the differential diagnosis of dementia between AD, FTD, VD, DLB and healthy controls using MRI-derived features from 504 patients. A number of features were extracted and combined using a Disease State Index (DSI) classifier [Mattila et al., 2012], including: brain structure volumes, voxel and tensor-based morphometry p-values for each structure, manifold learning projections, ROI-based grading features and the authors' own vascular burden measure. An overall balanced accuracy of 68.5% was achieved. These results were further improved upon in [Tong et al., 2017].

Further to this, data acquired from 117 patients at the Kuopio University Hospital was analysed in [Ángel Muñoz-Ruiz et al., 2016]. The DSI was used as part of a decision support tool to perform the differential diagnosis of AD and FTD based on clinical, neuropsychological, genetic, MRI, SPECT and CSF protein features and achieved a Area Under the Curve (AUC) of 0.97 and 0.94 for autopsy confirmed and clinically diagnosed cases respectively. Fifty subjects from the same cohort were further investigated in [Cajanus et al., 2018] with the aim of exploring the use of the same MR derived features as used in [Koikkalainen et al., 2016] above for the purpose of bvFTD differentiation from AD, DLB and healthy control subjects. They found that, while the features were useful for separating cases of bvFTD, AD and controls, the physical changes in bvFTD and DLB were too similar to be reliably differentiated.

2.6 Datasets

The work in this thesis is primarily technical, with novel techniques for analysing imaging data being proposed. A number of different datasets are therefore used throughout the thesis for their development and analysis. This section provides a brief description of each of these.

2.6.1 ADNI

The ADNI dataset provides a large amount of data for each patient including: clinical data (demographics, clinical and cognitive assessments), MR images (processed and unprocessed T_1- and T_2- weighted images, fMRI and Diffusion Tensor Imaging (DTI) available in the later studies), PET images (multiple tracers, processed and unprocessed PIB, FDG and Florbetapir), proteomic analysis and genotyping results. ADNI data is used in Chapters 7, 8 and 9.

2.6.2 Edinburgh SVD dataset

An SVD dataset was developed by colleagues at Edinburgh University and used extensively within this thesis. This dataset contains 147 fully annotated subjects suffering from SVD. It is a heterogeneous dataset containing data acquired using three different acquisition protocols. All image data were 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. Details of the protocols used for acquiring the data are given in Table 2.2, and their rationale is explained in [Valdes Hernandez et al., 2015]. Formal written consent from all subjects and ethical approval was acquired from the Lothian Research Ethics Committee (09/S1101/54, LREC/2003/2/29, REC 09/81101/54), the NHS Lothian R+D Office (2009/W/NEU/14), and the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and conducted according to the principles expressed in the Declaration of Helsinki.

[2] www.sbirc.ed.ac.uk

Table 2.2: Summary of the acquisition and segmentation protocols present in the Edinburgh dataset. [1][Valdes Hernandez et al., 2015, Valdés Hernández et al., 2013]

Protocol	1	2	3
Number	23	81	43
T_1 TR/TE/TI (ms)	9/440	9.7/3.984/500	
FLAIR TR/TE/ TI (ms)	9002/147/2200	9000/140/2200	
Ground Truth	Expert corrected histogram segmentation	Multispectral colour-fusion-based semi-automatic segmentation [1]	Expert corrected histogram segmentation
Lesion Types Present	WMH_{pvo}	WMH / Cortical infarcts	WMH_{pvo}

All image data were co-registered using FSL-FLIRT [Jenkinson et al., 2012] and mapped to the patients T_2-weighted space. Up to three lesions masks were created for each subject: WMH_{pvo}, and old and new stroke lesions. Lesions from images acquired under protocols 1 and 3 (Table 2.2) were extracted using histogram-based thresholding on FLAIR and manually rectified by an expert. Lesions from images acquired under protocol 2 were segmented by an expert following the procedure described in [Valdes Hernandez et al., 2015, Valdés Hernández et al., 2013], which uses a multispectral colour-fusion-based semi-automatic segmentation method and considers hyperintense signals that simultaneously appear in all T_2-based sequences.

As well as imaging data, clinical data including demographics, co-morbidities and imaging features were also made available. These include: age, gender, diabetes, hypertension, hyperlipidaemia, smoking, cholesterol, the number of perivascular spaces in the basal ganglia, and a measure of tissue atrophy. Imaging data from this dataset is used in Chapters 4 and 5, while the clinical data is also used in Chapter 4.

2.6.3 OASIS

The Open Access Series of Image Studies (OASIS) is a project aimed at compiling and freely distributing neuroimaging datasets in order to facilitate further research and discoveries in neuroscience. The first dataset, OASIS-1, contains cross-sectional MRI data for 416 subjects

aged 18-96, 100 of whom have been diagnosed with a stage of AD. OASIS-2 contains 150 older (60-97) subjects which were examined over two or more visits. 72 subjects remained non-demented throughout the study, 64 contained subjects which were, and remained, demented, while 14 developed dementia during the course of the study. OASIS-3 contains a retrospective dataset of 1098 patients compiled from across several studies over 30 years, including 609 non-demented adults, and 489 at various stages of cognitive decline. This dataset contains clinical data as well as data from a wide variety of MR and PET protocols. Data from OASIS-1 is used in Chapter 6.

Chapter 3

Technical Background

This chapter reviews and summarises some of the key areas of machine learning and computer vision explored in this thesis, particularly image synthesis, lesion segmentation and data augmentation. We start by reviewing the image synthesis techniques currently available, starting with purely generative methods for producing images from scratch, before moving on to conditional methods where images are generated from a given source image. As well as the methods themselves, we discuss their strengths and limitations, their similarities and differences, their applications, and potential avenues for further work. We then turn our attention to lesion segmentation and the various tools currently available, giving an overview of their development and function. Finally, we examine the role of data augmentation in neuroimaging. The purpose of this chapter is to both allow non-specialists to gain an overview of the field and to provide a library of information to expand on the concepts mentioned throughout this thesis.

3.1 Image Synthesis

Image synthesis is the process of producing an artificial image with a particular desired set of statistical properties. In medical image computing, this usually refers to generating images which appear to have been generated using a particular imaging modality. Methods to do this can be split into two broad categories: generative methods and conditional methods.

The goal of generative methods is to be able to generate realistic images from a target modality without a source image. These methods involve learning a low dimensional representation which can then be sampled and mapped to an image.

In conditional methods, often referred to as modality transformation methods in the context of medical imaging, the aim is to produce an image of one type, given an image of a different type. For example, generating a subject-specific CT image from an MR image for the purpose of attenuation correction in the reconstruction of a PET image acquired using a PET/MR scanner [Cardoso et al., 2015].

This section provides a review of each family of methods, including the key developments and applications in the field of medical imaging. We first explore the pure generative approaches, before reviewing methods for modality transformation.

3.1.1 Generative methods

In machine learning, a generative model is a model which allows for samples from a distribution to be generated once a set of underlying (latent) parameters are learned. An example of this is a Gaussian mixture model, where an unknown target distribution is considered to be able to be described as a weighted sum of K Gaussian distributions. The probability of a particular x value being from this distribution can be given by,

$$p(x) = \sum_{i=1}^{K} \phi_i \mathcal{N}(\mu_i, \theta_i), \tag{3.1}$$

where ϕ_i, μ_i and θ_i are respectively the weight, mean and standard deviation of the i^{th} component of the model. The Expectation Maximisation (EM) [Dempster et al., 1977] algorithm can be used to estimate the value of these latent variables given a sufficient number of real examples from the unknown distribution. New unseen samples can then be generated by sampling from the learned distributions. These types of approaches work well in low dimensional cases and where the expected form of the distribution can be predicted, however medical images are in-

herently high dimensional with a typical MR image containing over 1 million voxels, and their distribution within this high dimensional space is hard to predict. Techniques such as EM are underdetermined when there are more parameters to learn than available examples and would, therefore, require some form of dimensionality reduction or regularisation to be used on medical images.

Deep learning has provided an alternative, data-driven, approach to generative models in autoencoders and Generative Adversarial Networks (GANs). Both approaches make no assumption on the inherent structure within the data and instead provide a mapping from a simple (usually a multi-dimensional Gaussian or uniform) distribution directly to the high dimensional image distribution through a series of non-linear transformations. These transformations take the form of a deep neural network which is learned using real training examples. Once the model has been learned, new images can be generated by sampling from the simple distribution and passing the outputs through the learned network.

Autoencoders

Autoencoders were first proposed in [Vincent et al., 2010]. The aim of an autoencoder is to learn a low dimensional representation of a distribution by training on examples from this distribution. In the case of large images, this involves chaining two Convolutional Neural Networks (CNNs) together: an encoder and a decoder. The task of the encoder is to take as input an image and to output a lower dimensionality encoding. The decoder's role is to map this encoding back into the original image with as little lost information as possible. Training an autoencoder is simple and done in an end-to-end fashion. Real images are fed into the input of the encoder network and passed through both networks, with a loss calculated between the output of the decoder and the original image. The networks are optimised by minimising this loss.

In order for the image to be accurately reconstructed by the decoder, the encoding provided by the encoder must contain all the relevant information to describe any given image from the distribution of training samples. This provides a low dimensional latent representation, which,

combined with the decoder, yields a model for the image distribution. The size of this latent space can be tuned empirically to balance the amount of compression provided with the quality of the reconstructed images, with smaller latent spaces leading to greater compression at the cost of a higher reconstruction error.

However, this formulation does not provide a complete generative model as there are no constraints on the distribution of the latent representation. It is therefore not possible to sample unseen images using the model since it is not possible to know what is a valid latent encoding. As a solution to this, Variational Autoencoders (VAEs) were proposed [Kingma and Welling, 2013] to constrain the latent representation to form a Gaussian distribution. Sampling an unseen image is then simply a matter of sampling from a Gaussian distribution and passing this through the decoder.

Generative Adversarial Networks

An alternative to VAEs is provided by GANs. First proposed in [Goodfellow et al., 2014] GANs are a class of neural networks which aim to learn to produce samples from a given distribution. The task is formulated as a game with two players. The first player, the generator, attempts to produce samples which the second player, the discriminator, cannot identify as fake when compared to a set of real examples. A common analogy is that of an art forger (generator) and a detective (discriminator). The forger has never seen a portrait before, but wants to be able to paint them. The detective has no knowledge of art, but has a set of example portraits. The game proceeds to cycle between the two players, with the forger attempting to paint a portrait, and the detective attempting to identify the forgery by comparing it to real portraits. As the game progresses, the forger learns what sort of features trick the detective and uses more of these, leading the detective to have to become more sophisticated in turn. The two players will continue to get better at their job, driving the other to improve. When the game ends, the forger is able to produce highly detailed portraits, despite never having seen one, while the detective is now an expert at spotting forged portraits.

In practice, these roles are taken by two neural networks. The generator maps a random vector

z, which forms the latent representation, to an image, while the discriminator takes an image and outputs a belief as to whether it's real or synthetic. During training, the discriminator is given a batch of real and synthetic images to learn from, with the loss function being a form of classification error. The generator then produces a new batch of synthetic images to test the discriminator, with its loss function being inversely proportional to the discriminator's loss on this batch.

The original GAN formulation as proposed in [Goodfellow et al., 2014] involves a generator network G and discriminator network D, updated using the following algorithm:

while *stopping criteria not reached* **do**
  for *k steps* **do**
    Sample batch Z of size n from noise distribution
    Compute $E^G = \sum_{i=1}^{n} \mathrm{D}(\mathrm{G}(Z_i))$
    Sample batch X of size n from real images
    Compute $E^X = \sum_{i=1}^{n} \mathrm{D}(X_i)$
    Update parameters of D, θ_d, by ascending $\Delta_{\theta_d} \frac{1}{n} \sum_{n}^{i=1} [log(E_i^X) + log(1 - E_i^G)]$
  end
  Sample new batch Z of size n from noise distribution
  Compute $E^G = \sum_{i=1}^{n} \mathrm{D}(\mathrm{G}(Z_i))$
  Update parameters of G, θ_g, by descending $\Delta_{\theta_g} \frac{1}{n} \sum_{n}^{i=1} log(1 - E_i^G)$
end

Algorithm 1: GAN training algorithm proposed in [Goodfellow et al., 2014].

where n is the batch size used during each iteration, and k is the number of times the discriminator is updated for each time the generator updates, set to 1 and in [Goodfellow et al., 2014].

Many approaches have since been developed based upon this framework, as well as several offshoots such as conditional GANs which introduced adversarial losses to tasks such as style transfer and super-resolution. These will be revisited later in Section 3.1.2. First, the developments to this basic GAN formulation will be discussed, maintaining the focus on pure generative models.

An early development came with Laplacian GANs [Denton et al., 2015]. Here, the authors proposed a multi-resolution approach, with a series of generator/discriminator pairs trained to produce difference images leading to successively higher resolution images. During sampling, the first generator produces a low resolution (8-by-8px) image from a random input vector

$\mathbf{z_3}$. The resulting image is upsampled to 16-by-16px and fed as input into the next generator along with a new random vector $\mathbf{z_2}$. The output of this generator is a difference image which, when added to the upsampled input, produces a realistic higher resolution image. This process is repeated twice more resulting in an image of size 64-by-64px. While this method is able to produce higher quality larger images than the original GAN, the requirement for multiple GANs means longer training times, and the inclusion of conditional elements in the sampling procedure means significantly longer sampling times. There has been little further work in this direction, however, the ideas of using multiple resolutions and performing conditional GAN based super-resolution were developed further in [Karras et al., 2017] and [Ledig et al., 2016] respectively with great success.

A major improvement was made on the original GAN in [Radford et al., 2015] with Deep Convolutional Generative Adversarial Networks (DCGANs) where training stability and image size and quality were improved by modifying the architecture of D and G to remove fully connected layers and replace pooling layers with (fractional) strided convolutions. The authors also investigate the latent representation provided by the input vector $\mathbf{z}$. They demonstrate that interpolation through this space from one vector to another results in a semantically sensible transition from one image to the other in image space. Objects present in one image gradually transform into objects present in the other, with each image in the transition appearing plausible. They also show that arithmetic in this latent space is also semantically reasonable in image space. They first compute the average $\mathbf{z}$ vectors which correspond to images of smiling female faces ($\bar{\mathbf{z}}_{fs}$), neutral female faces ($\bar{\mathbf{z}}_{fn}$) and neutral male faces ($\bar{\mathbf{z}}_{mn}$). They then show that if they take the vector $\hat{\mathbf{z}} = \bar{\mathbf{z}}_{fs} - \bar{\mathbf{z}}_{fn} + \bar{\mathbf{z}}_{mn}$ the result in image space would be that of a smiling male face - a semantically sensible outcome. A number of incremental improvements were made in [Zhang et al., 2018] resulting in more stable training at higher resolutions.

Further developments upon this framework included using a least squared loss for the discriminator [Mao et al., 2017], while another major step forward was made in [Arjovsky et al., 2017] where the authors backed up some convincing theoretical proposals with strong practical results. One of the criticisms of the previous GAN formulations was the lack of a true image quality related cost function, as well as there being a need to balance the number of training cycles

provided to the two networks. Both of these problems were addressed in [Arjovsky et al., 2017] by proposing the Wasserstein Generative Adversarial Network (WGAN). Their main contribution was to replace the Jenson-Shanon divergence approximating formulation of the previous work, with one in which the discriminator approximates the Wasserstein distance instead. The Wasserstein distance is an earth mover distance, which measures the total amount of work which would need to be done to transform one distribution into the other by moving "mass" from one region of the distribution to another region. This involves relatively simple changes in the update steps in Algorithm 1: $\Delta_{\theta_d} \frac{1}{n} \sum_{n}^{i=1} [E_i^R - E_i^G]$ and $\Delta_{\theta_g} \frac{1}{n} \sum_{n}^{i=1} E_i^G$ for D and G respectively. However, this solution to compute the Wasserstein distance is only valid over functions which are 1-Lipschitz, ie, there are no two points on the function which are connected by a slope with a gradient greater than 1. One final step is therefore required for the discriminator to be a reasonable estimator of the Wasserstein distance, which is to constrain the weights to lie within a fixed box. This is performed by simply clipping every weight to the range $[-0.01, 0.01]$ after each update step. The authors admit that this is not an elegant solution, however, the resulting GAN proves to be very stable and effective on a wide variety of tasks. In addition, with the optimally trained discriminator now providing an approximation of the Wasserstein distance, the formulation has an image quality related cost function.

The issue of weight clipping was addressed in [Gulrajani et al., 2017] with the clipping procedure replaced by an additional gradient norm penalty term within the discriminator objective function. This naturally encourages the discriminator towards solutions which are 1-Lipschitz. They demonstrate that the weight clipping approach leads to much simpler functions being learned and that by replacing this with the additional penalty term, the capacity of the discriminator increases, allowing for higher moments of the distribution to be learned.

A family of GANs which utilise an auto-encoder as the discriminator have also been proposed [Zhao et al., 2016, Berthelot et al., 2017]. The theory is that a network which has been trained to encode and decode a real image with minimal loss should also be able to do the same to a synthetic image. Rather than training a generator to minimise the discriminator loss, the generator is trained to match auto-encoder loss distributions between the real and synthetic images.

Another significant step towards larger image generation came with the Progressive Growing of GANs (PGGAN) [Karras et al., 2017]. The proposed multi-resolution approach to training is shown to reliably generate images up to 1024-by-1024px, well beyond the maximum of 256-by-256px seen in previous formulations. It does this by progressively growing the size of the networks, starting with a small GAN generating 4-by-4px images, and adding layers to both the generator and discriminator throughout training, successively doubling the output image size until the desired level is reached.

Within the PGGAN framework, the architecture follows the standard decoder- (generator) encoder (discriminator) pattern, with the generator mapping from a latent vector to an image, and discriminator mapping from an image to a single number. With an input vector of size $|\mathbf{z}|$, an initial 4-by-4 convolutional layer is followed by a 3-by-3 convolutional layer, each with $|\mathbf{z}|/2$ filters. These are then followed by a series of blocks, each consisting of a nearest-neighbour up-sampling layer and two 3-by-3 convolutional layers, with each layer in a block containing half the number of filters as those in the previous block. Once the feature map size reaches the desired image size, a final 1-by-1 convolutional layer compresses these to the final output image (typically with 3 channels, for RGB-images). The discriminator architecture reflects the generator, with down-sampling layers in place of the up-sampling layers, and the final layer mapping to a single output value.

During training, the constituent up/down-sample blocks are introduced over time. The initial architecture generates 4-by-4px images, with this size increasing as more blocks are added. The transitions from one image size to the next within the generator are performed gradually, with the output image during a transition being a weighted average between the up-sampled output from the previous block, and the output of the new block. This weighting is changed linearly through the transition period, thereby gradually increasing the impact of the new block. The same processes are mirrored in the discriminator. In this way, the effect of suddenly introducing randomly initialised parameters into the network is minimised, thereby maintaining consistent image generation. Transition periods are measured in number of iterations and are typically equal to the number of iterations spent between each transition.

Developments in GAN architectures and training have been primarily driven by the computer vision community for tasks such as natural image and video generation, texture synthesis, image inpainting, style transfer and image editing. A discussion of these applications is beyond the scope of this thesis, however, a comprehensive and accessible review of these applications up to the end of 2017 is available in [Wu et al., 2017]. Applications of GANs in medical imaging are reviewed further in the context of modality transformation in the next section, whilst the use of GANs in data augmentation is reviewed at the end of this chapter.

Discussion

One of the main challenges in developing generative models is that of evaluation. Beyond visual inspection, it is often difficult to assess the quality of generated images. However, a few metrics have been proposed in the context of GANs to compare properties of the distribution of the generated images to that of the real images:

- *Wasserstein distance*

  As described previously, the Wasserstein distance can be used to measure the similarity between two distributions, with [Arjovsky et al., 2017] showing that this could provide an optimisable cost function which can be learned by a network. However, this only provides an estimate, which may vary significantly between architectures and training times. It is therefore not always suitable when comparing the outputs of different networks.

- *Multi-scale structural similarity (MS-SSIM)*

  MS-SSIM was first proposed in [Odena et al., 2016] for use in GANs as a measure of image diversity, exploiting an approach developed previously [Wang et al., 2004] for measuring perceived image quality. It provides a score between 0 and 1, with higher values corresponding to perceptually more similar images. A metric for image diversity can be computed by calculating the mean score between randomly selected synthetic images, with a lower score indicating less similar images and therefore more diversity. By comparing this mean score to one calculated on the training images, the diversity of the

generated images can be compared to that of the training images. Note that this does not assess image quality, only diversity.

- *Sliced Wasserstein Distance (SWD)*

  Proposed in [Karras et al., 2017], the SWD is another multi-scale approach which aims to assess both image diversity and quality revolving around extracting patches at multiple resolution levels and computing the similarity between the patches extracted from the real and synthetic images at each level. A set of images is first sampled. The Laplacian pyramid for each image is computed by first downsampling the image to a base resolution. Each successive level is defined as the difference image between an image downsampled to twice the resolution as the previous level, and a 2-times upsampled version of the previous level. From each of these levels, a set of patches are randomly sampled to form a descriptor. These are grouped together with other descriptors extracted from the same level from other images and are normalised (within channels if the image is multi-channel). Once all the descriptors are extracted, the sliced Wasserstein distance (an efficient approximation to the Wasserstein distance) is computed between the descriptors extracted from the real and synthetic images at each resolution level. This approach allows for a detailed understanding of where the two sets of images are different or similar. A low score implies a high degree of similarity, so a low score at the base resolution level suggests similarity within the low-frequency features such as broad shapes, wheres low scores at higher resolutions suggest similarity at higher frequencies such as the fine details and edges.

  The sliced Wasserstein distance, as defined in [Rabin et al., 2012] estimates the Wasserstein distance by repeatedly projecting the distribution of descriptors onto a series of 1D spaces. The Wasserstein distance in these spaces can then be calculated trivially and summed to provide a score.

- *Inception score*

  Used in [Zhang et al., 2018], the inception score was calculated by using an additional pre-trained model to analyse the generated images to produce a conditional label dis-

tribution. The motivation was that meaningful images will result in individual label distributions with low entropy, while varied images will produce a marginal across a set of generated images with high entropy. By combining these two, a score can be produced which correlates well with visual quality and human classification. This is a useful tool for developing pure generative GAN models and testing modifications and different techniques. However, in order to be meaningful, it can only be used to compare methods which have been trained on the same dataset, and that dataset must also be related to the dataset used in training the classification model (ImageNet in this case). This limits its applicability in niche fields such as medical imaging.

- *Human testing*

  Given that the original aim of GANs was to be able to generate images which are indistinguishable from real images to the human eye, it is perhaps unsurprising that the most realistic method for evaluation is to ask humans to try and identify whether images are real or synthetic. The authors of [Zhang et al., 2018] used Amazon's Mechanical Turk (a crowdsourcing platform which allows users to be paid small amounts to perform simple web-based tasks) to ask people to differentiate between real and synthetic images. This provided some useful scores for comparing methods. However, they found that results from experiments within their own group differed significantly from those provided by the users of Mechanical Turk. This raises questions over whether such a platform is a reliable method of evaluation. A similar approach was taken in [Chuquicusma et al., 2018] where the authors asked two radiologists to identify synthetic lung nodules. Whilst they could consistently fool the more junior radiologist with 4 years experience, the senior radiologist with 13 years experience could identify the synthetic lesions much more frequently.

  Both of these results demonstrate the challenges of using humans to analyse synthetic images. Experience, motivation and exposure to the type of images being analysed will all play a part in the resulting scores. Being able to get meaningful results to compare across experiments is challenging, and across studies, nearly impossible.

As it stands, the gold standard for evaluating a novel pure generative GAN technique or ar-

chitecture is the inception score. Whilst allowing for meaningful comparisons between other methods and the GAN itself, it cannot be used for evaluating performance on different datasets. In these cases, a combination of MS-SSIM and SWD provides a good alternative, allowing comparisons to be made between any two sets of images. Using humans for analysis, whilst arguably providing the measure most suitable to the task, has many complications which prevent such a measure being used for anything more than a broad analysis.

There are, however, alternatives. GANs are often used as part of a bigger system or for a particular application, such as domain adaptation, semantic inpainting and super-resolution. In these cases, the best way to evaluate the GAN is simply to evaluate the system as a whole. Changing the architecture or training procedure of the GAN should have a measurable impact on the application the GAN is being used for, allowing for the most suitable architecture and hyper-parameters to be found. This can, however, be an extremely slow process, especially if the application requires long training times itself. In addition, care must be taken to avoid overfitting a single dataset when taking this approach to evaluation. It is important to ensure that standard rules regarding the separation of a test, validation and training set are followed for the entire system, not only its individual components. For example, data which is used to train the GAN component must not later form part of the test set for the entire system.

While the above approaches to evaluation look to measure the ability of methods to perform pure image generation, it is also worth considering how to evaluate conditional GANs, where one image is transformed into another image usually with the same content but a different appearance. The above metrics which measure variation and "perceptual similarity" do not take into account the need for the generated image to contain the underlying information embedded in the original image. In these cases, it is necessary to include an additional metric such as the Euclidean distance, cross correlation or mutual information between the two images. By doing this, it is possible to evaluate an image based upon two measures. The first asks "Does the generated image look like it belongs in the class of images to which it is being transformed?", while the second asks "Does the generated image have the same content as the original image?".

In Chapters 5 and 6 of this thesis, GANs are used primarily as components within a larger

system. We, therefore, perform evaluation by measuring the effect of the GAN on the system as a whole. In Chapters 7 and 8 we train GANs using a Wasserstein loss and use this to assess convergence prior to using the trained models.

3.1.2 Modality transformation

Early work [Hertzmann et al., 2001] in the field of natural image processing introduced the idea of modality transformation under the moniker Image Analogies (IA). This work brought together and extended a number of previous pieces of work from across the fields of machine learning and image processing to present a flexible framework to perform image synthesis as an effective solution to a variety of general image processing problems. These include super-resolution, where the aim is to synthesise a higher resolution copy of an image; texture transfer, where a source texture is applied to a new image; and artistic filters, where images with the appearance of being produced using particular artistic styles are produced from images with different artistic styles.

The idea was picked up by the medical imaging community as a tool to help solve a number of problems being faced in the study of medical images. This work can be mostly split into two main fields, super-resolution and modality transformation. The latter drawing parallels with the artistic filters described in [Hertzmann et al., 2001] in being the task of generating an image with the appearance of one imaging modality by using information from one or more images from other modalities.

The ability to do this accurately allows for a number of interesting applications, for example, in multi-modality registration [Iglesias et al., 2013, Dawant et al., 2012, Cao et al., 2014, Cao et al., 2013, Roy et al., 2014a, Kroon and Slump, 2009, Chen et al., 2015b] where the problem can be reduced to mono-modality registration when one modality is synthesised from the other. There are also algorithms for segmentation or classification which require an input image from a certain modality. The ability to synthesise these modalities from another modality can expand the applicability of these algorithms, as demonstrated in [Roy et al., 2010, Roy et al., 2013, Jog et al., 2015]. More recently, as alluded to earlier, the ability to synthesise

a CT image from an MR image has received particular attention [Cardoso et al., 2015] as a result of the rise of PET/MR scanners. An attenuation map is required to accurately reconstruct a PET image. This is readily available in the form of a CT image in PET/CT systems, however it must be inferred from the MR in PET/MR systems. Being able to synthesise an accurate CT from an MR image solves this problem.

3.1.3 Overview of methods

This section contains an overview of the image synthesis methods currently available in the literature. Whilst being closely coupled with the field of super-resolution, the focus of this survey is to provide a background to modality transformation in the field of MRI. However, key papers containing important developments from other fields including super-resolution and non-MR medical image synthesis which have been transferred to the problem of MR modality transformation have also been included. Papers pertaining to non-pathological image synthesis have also been included, regardless of the modalities involved.

The literature is presented in four groups. First, are those methods which are derived from the IA approach mentioned earlier. Next, are a set of methods which perform synthesis by using regression techniques. Non-deep learning approaches which don't fall into the previous two categories then are presented, with deep learning methods reviewed finally.

Patch matching approaches

The framework proposed in [Hertzmann et al., 2001], which forms the foundation of many of the algorithms developed since, uses a source image B along with co-registered images A and A', where A and B share the same statistical properties. The aim is then to synthesise B' such that B' is related to B in the same way A' is related to A. The basic framework proposed employs a patch based approach, whereby for each patch in B, the closest patch A is found, and the corresponding patch in A' is propagated to the same location in output image B'. This framework asks one major design question, the development of which has been the focus of

much of the later work.

The question relates to finding the best match in A for a patch in B. Whilst A and A' are a single image pair, they can be thought of more simply as a collection of many pairs of corresponding patches. There is, therefore, no reason to limit A and A' to being a single image pair when there may be more examples $\{A_1, A'_1; A_2, A'_2...\}$ which also share the same statistical relationship. In the case where many image pairs exist, there can be a very large number of patch pairs. Exhaustively searching this large library for the optimum match for a given patch from B can, therefore, be extremely time-consuming. Any method proposed based upon this framework must, therefore, address this problem and find a balance between finding the most similar patch, and the speed of computation. A related problem is that of deciding what similarity metric to use to decide on the closest match. As pointed out in [Hertzmann et al., 2001], a simple L_2-norm may not necessarily provide the greatest perceptual similarity, and an alternative may provide better, if slower, results.

Another important issue raised in [Hertzmann et al., 2001] is that of luminance remapping, or intensity normalisation. If intensities in image B and in the library of patches $\{A, A'\}$ do not have an exact correspondence, then the search for an appropriate match may yield a suboptimal result. This is a particular problem when using an L_2-norm for similarity. This problem can be demonstrated by considering A to be simply a darker version of B. Searching for the closest patch in A to a given patch in B using an L_2-norm is unlikely to result in the correct patch being found due to the bulk intensity difference.

In [Hertzmann et al., 2001], the solution to the question of speed is to employ a multi-scale approach using Gaussian pyramids and an approximate nearest neighbour approach to finding the closest match. This method also benefits from using a relatively small library $\{A, A'\}$ comprising of only one image. The L_2-norm is used for searching, however, in order to enforce spacial coherency, the distances found are scaled by a factor which takes into account how coherent a given patch will be given any already synthesised surrounding pixels. To address the problem of bulk intensity differences, the authors propose applying a linear scaling to both A and A' which matches the means and variances of A to that of B.

Several more proposals have been made based upon the IA framework which address these problems in different ways. In [Roy et al., 2010], the authors apply the framework to the problem of FLAIR image synthesis by augmenting source B and library $\{A, A'\}$ to contain information from T_1- and T_2-weighted (T_1 and T_2) images, $\{B_{T1}, B_{T2}\}$ and $\{A_{T1}, A_{T2}, A'\}$ respectively. The similarity metric used being the weighted sum of the L_2-norms between corresponding patches from both modalities and a coherence factor, seen in Equation 3.2.

$$\begin{aligned} Distance(i,j) =& [w_{T1}(j)||B_{T1}(i) - A_{T1}(j)||^2 \\ &+ w_{T2}(j)||B_{T2}(i) - A_{T2}(j)||^2 \\ &+ \lambda R(B_{T1}(i), A_{T1}(j), B_{T2}(i), A_{T2}(j))], \end{aligned} \tag{3.2}$$

where i and j are the location of patches in the source and library images respectively, w_{T1} and w_{T2} are spatially varying weights, λ is a weight controlling the relative importance of coherency, and R is the coherence function shown in Equation 3.3.

$$\begin{aligned} R(B_{T1}(i), A_{T1}(j), B_{T2}(i), A_{T2}(j)) = \sum_{k \in N_i} \sum_{l \in N_i} [&||B_{T1}(k) - A_{T1}(l)||^2 \\ &+ ||B_{T2}(k) - A_{T2}(l)||^2], \end{aligned} \tag{3.3}$$

where N_i and N_j are the neighbourhoods of i and j.

The authors add an additional step to the process by combining a number of the most similar patches found using a non-local means [Buades et al., 2005] method. In [Roy et al., 2014b], the authors use the same method to synthesise higher resolution versions of B' by blurring and under-sampling $\{B_{T1}, B_{T2}\}$ and $\{A_{T1}, A_{T2}\}$. In both papers, the authors make the key assumption that B_{T1} and A_{T1}, and, A_{T2} and B_{T2} have the same respective intensity distributions and therefore do not require normalisation to allow for the use of the L_2-norm. They also propose no method of speeding up the search, instead relying on a relatively small patch library derived from a single image triplet.

This method was further evolved to produce the publicly available[1] MIMECS tool, presented in [Roy et al., 2011] as a means to normalise images produced through different T_1 acquisition protocols to a standard intensity space in order to achieve more consistent anatomical segmentations. The problem reverts back to one of synthesising an image B' from corresponding source image B and library $\{A, A'\}$. The non-local means method for combining patches and coherence constraint is replaced by a method which uses the closest patches to regularise the ill-posed inverse of the equation governing the physical acquisition of MR images, using techniques from compressed sensing [Donoho, 2006]. The problem of search speed is addressed by limiting the search space of A to locations which have same tissue type as the patch taken from B, by applying a rough tissue segmentation algorithm to divide A and B into 6 tissue types. Finally, the issue of intensity normalisation prior to synthesis is addressed by applying a linear scaling to ensure that the peak of the histogram corresponding to the intensity of the white matter is 1 in all images.

A large scale analysis of the applicability of image synthesis using MIMECS was presented in [Roy et al., 2013]. The authors demonstrate the use of image synthesis in longitudinal data normalisation, atlas construction, contrast normalisation (as in [Roy et al., 2011]), distortion correction in diffusion images, super-resolution and FLAIR image synthesis. During the evaluation of the latter, the authors identify a limitation of MIMECS when it comes to its ability to synthesise lesions. They suggest that this inability to synthesise lesions could be seen as a useful feature for the purpose of lesion segmentation, through the subtraction of the synthetic FLAIR image from the real FLAIR image, though no investigation into this is carried out.

Another technique to speed up the search for patches was introduced to the basic framework in [Iglesias et al., 2013]. Using a patch library derived from multiple images, the search space of A was reduced by first co-registering the images and then restricting the search to patches coming from the areas close to the source patch in B. This, along with a hierarchical approach, allowed for a patch library taken from 39 images to be searched in a reasonable time. This approach was also used in [Konukoglu et al., 2013] where a Bayesian approach aims to find the

[1] www.nitrc.org/projects/image_synthesis

most likely patch from the library to use for the purpose of super-resolution.

Sparse coding [Aharon et al., 2006] was suggested as an efficient method to reconstruct patches from a library for the purpose of super-resolution in [Rueda et al., 2013], which was further developed in [Huang and Wang, 2013]. The motivation behind this approach relates to how the optimal reconstruction of a patch to be used in B' will be a linear combination of patches in A'. The coefficients of this combination will be found be seeing which combination of patches in A best reconstruct B. However, a sparse solution is desired as the correct result is more likely to require a few similar patches to be combined, rather than a large number of dissimilar patches. Such a sparse representation of a patch taken from location i in source B can be found by solving,

$$\alpha^* = \arg\min_{\alpha}[\lambda||\alpha||_1 + \frac{1}{2}||A\alpha - B(i)||_2^2], \tag{3.4}$$

and the synthetic patch can be found through $A'\alpha^*$.

Since this approach treats each patch irrespective of its neighbours, a further global regularisation step is applied to enforce continuity. However, this step uses information from the original low-resolution image B, exploiting the fact they are from the same modality. Such an approach may not be possible in the case of modality transformation.

Sparse coding was also employed for the registration of microscopy images using an otherwise standard IA framework, first in [Dawant et al., 2012] and evaluated further in [Cao et al., 2014]. These propose novel solutions to the problem of searching the dictionary and intensity normalisation. The former is handled by randomly selecting a subset of the full dictionary to search, whilst the latter is addressed by a simple scaling of image intensities to the range $[0, 1]$.

An alternative approach to the data based global regularisation used in [Rueda et al., 2013] to ensure coherence was proposed in [Ye et al., 2013]. This paper looks at the synthesis of T_2 and diffusion tensor imaging fractional anisotropy maps from T_1 images, and introduces the idea of the deliberate (as opposed to being observed as an unintended consequence in [Roy et al., 2013])

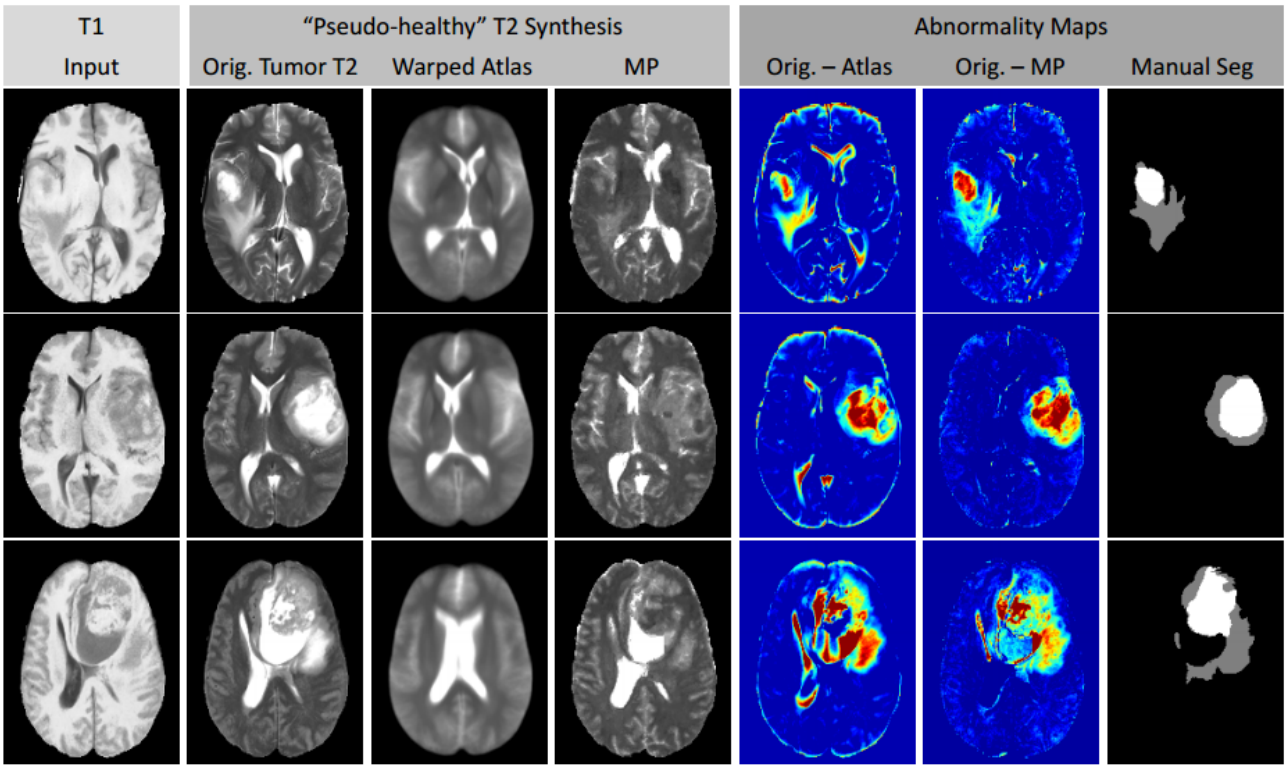


Figure 3.1: Three example cases where pseudo-healthy T_2 synthesis has been used to locate abnormalities, reproduced from [Ye et al., 2013] with permission. Warped atlas refers to the approach given in [Miller et al., 1993], whereas MP refers to Modality Propagation, the name given to the proposed method.

synthesis of a non-pathological image B' from a pathological source image B, or "pseudo-healthy" image synthesis as they term it. Whilst not performing any objective analysis, the authors show that synthesising a subject-specific pseudo-healthy image, and subtracting this from the acquired image, provides a visually much clearer indication of pathology than when subtracting a non-rigidly aligned atlas found using a similar approach to that described in [Miller et al., 1993], shown in Figure 3.1.

The proposed method uses the basic IA framework, with a number of alterations. Firstly, the library $\{A, A'\}$ is comprised of up to 100 image pairs, thereby being heavily reliant on a fast searching strategy. The authors use the local search approach used earlier [Iglesias et al., 2013, Konukoglu et al., 2013] to limit the search for matching patches to those which came from nearby locations through affine alignment. They also apply a novel approach whereby they reduce the size of the library by only searching patches which come from the most similar images, with similarity defined as the L_2-norm at a local level. Put explicitly, when synthesising a

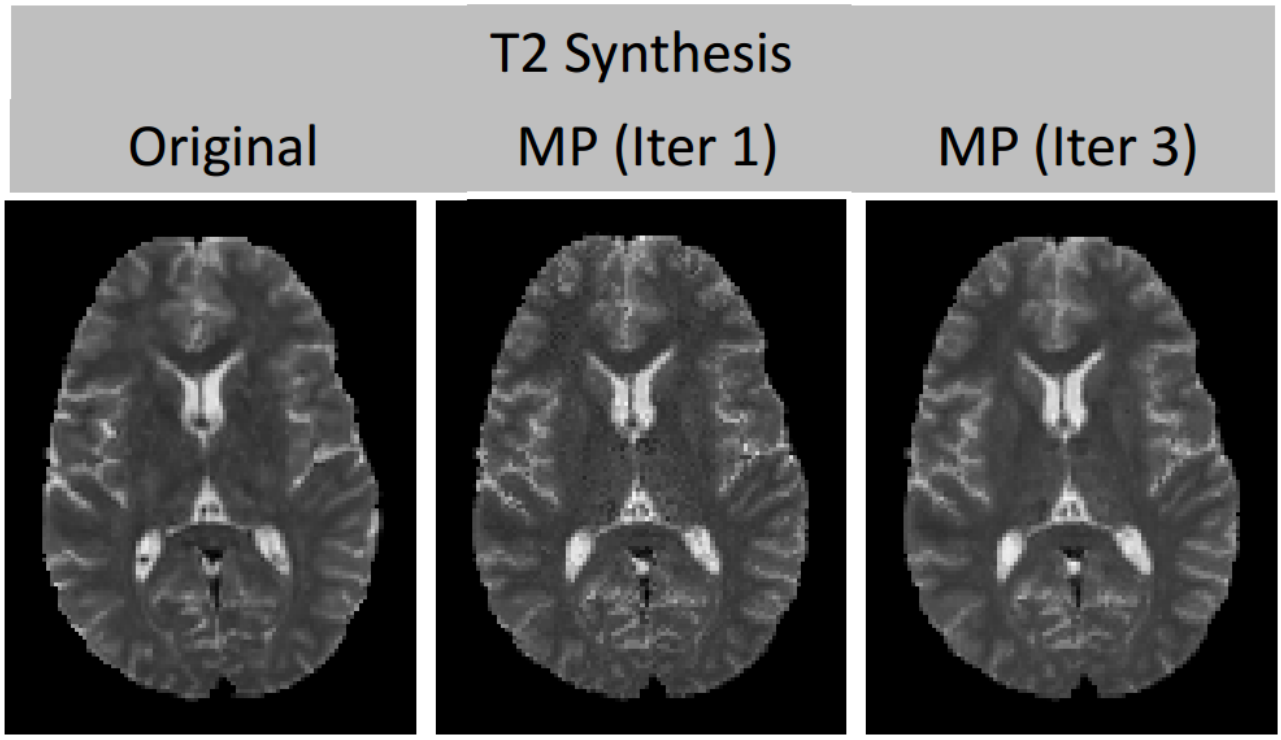


Figure 3.2: The benefits of the proposed iterative approach, reproduced from [Ye et al., 2013] with permission. The image corresponding to a single iteration clearly appears more noisy, with the result after 3 iterations displaying fine details much more clearly.

patch at position $\mathbf{x}$, the algorithm first looks at a large (approx. 3 times patch size) window around $\mathbf{x}$ in each image. The most similar k images are found and the patches which fall within this window in each image are searched to find the closest patch to the one at $\mathbf{x}$. The authors also employ an iterative approach to encourage coherency in the synthesised images. During the first iteration, the image is synthesised as described. In subsequent iterations, the L_2-norm cost function is augmented by the addition of a term to reflect the similarity of the chosen patch to the patch taken from the previously synthesised image. The relative weighting, $\alpha \in [0, 1]$, of these two terms can be controlled to balance the two effects (Equation 3.5). The authors demonstrate that this iterative approach improves synthesis results, demonstrated in Figure 3.2.

$$Distance(i, j) = (1 - \alpha)||B_{T1}(i) - A_{T1}(j)||^2 + \alpha||B'_{T2}(i) - A'_{T2}(j)||^2. \tag{3.5}$$

Another use of pseudo-healthy image synthesis was presented in [Tsunoda et al., 2014]. Here the authors aimed to synthesise pseudo-healthy chest radiographs for the purpose of lung nodule detection. The motivation behind this is the same as in [Ye et al., 2013]: the ability to synthesise a pathology free image allows for abnormalities to be detected by the subtraction of this image from the true acquired radiograph. The proposed algorithm optimises the search speed using the commonly used database alignment and local search method, as well as heavily (64×) downsampling the library images. To account for bulk intensity variations, the authors use normalised correlation coefficient R as a similarity metric (Equation 3.6), in place of the usual L_2-norm. Again, no objective measures or solid segmentations are calculated, however, the images provided clearly demonstrate the benefits of such a method.

$$R = \frac{1}{n}\sum_{x=1}^{n}\frac{(f(x)-\bar{f})(t(x)-\bar{t})}{\sigma_f\sigma_t}, \tag{3.6}$$

where $f(x)$ and $t(x)$ are the intensities at position x of two patches being compared, n is the number of voxels in each patch, $\bar{f}$ and $\bar{t}$ are the mean intensities of the two patches, and σ_f and σ_t are the standard deviations of the intensities in each patch.

A final family of methods based loosely on the original IA framework and incorporating Bayesian approaches have also been proposed [Konukoglu et al., 2013, Cao et al., 2013, Roy et al., 2014a]. These methods aim to find the probabilistically most similar patches in the library by treating a patch as being drawn from a Gaussian mixture model with unknown mean and variance. The goal is then to find the most likely patches in A given a patch in B. This can be solved iteratively by using the EM algorithm. In [Konukoglu et al., 2013], the authors claim the full algorithm to be intractable on MRI volumes and therefore propose an approximation. On the other hand, the authors of [Cao et al., 2013] use the full algorithm incorporating sparse coding and showed it to be applicable in the multi-modal registration of microscopy images. The authors of [Roy et al., 2014a] also use the full algorithm for the purposes of MR to CT synthesis for multi-modal registration. They point out that the algorithm has a complexity of $O(NL^2)$ with L being the number of patches in library $\{A, A'\}$. For large L, this clearly

becomes intractable, as such they limit L to the 40 nearest neighbours to make the complexity manageable.

A comparison of the methods described so far can be seen in Appendix B, including the approaches taken to tackle the common problems and their application.

Regression methods

This family of approaches to the problem of image synthesis revolve around finding a function which will map intensities from the source modality to those of the target modality. This function is learned from a set of pairs of co-registered training images and can be spatially variant.

The method proposed in [Kroon and Slump, 2009] to aid in registration aims to learn the correspondence between the intensities of T_1 and T_2 images by finding the most common intensity in the target modality for a given intensity in the source modality at each voxel, weighting the contribution of each surrounding voxel by it's distance using a Gaussian kernel. A single image pair is used to train, and due to the need to have a large number of samples in order to calculate a robust mode for each intensity value, large windows around each voxel are used. The authors do not say how many bins they use to calculate the modes or give any indication of any intensity normalisation procedure.

Put explicitly, to synthesis a voxel at location $\mathbf{x}$, the N bin joint histogram $\mathbf{H_x}$ of the two modalities around $\mathbf{x}$ must first be formed. Assuming A and A' to be scaled to the range $[0, 1]$, this can be achieved by iterating though all the voxels y in a window around $\mathbf{x}$,

$$\mathbf{H_x}(\lfloor A(y)N \rfloor, \lfloor A'(y)N \rfloor) = \mathbf{H_x}(\lfloor A(y)N \rfloor, \lfloor A'(y)N \rfloor) + e^{\frac{||y-x||^2}{2\sigma^2}}. \tag{3.7}$$

B' can subsequently be synthesised at voxel $\mathbf{x}$ as follows,

$$B'_x = \arg\max_j [\mathbf{H}_{\mathbf{x}}(\lfloor B(y)N \rfloor, \lfloor jN \rfloor)]. \tag{3.8}$$

Another regression method was proposed in [Jog et al., 2013] whereby the relationship between the intensities of patches from T_1 and T_2 images, and 1.5T T_1 and 3T T_1 images, are learnt through the use of a non- spatially variant regression forest. A single pair of atlas images are used for training. Whilst not providing the full specification of the machine being used, the authors do provide an interesting comparison to other methods in terms of computation time. They report their method takes 3 hours to train their model, and 5-10 minutes to synthesise, compared to approx. 2-3 hours for [Roy et al., 2011] and approx. 1 hour for [Miller et al., 1993]. They also provide objective metrics (Mean Squared Error (MSE) and Universal Quality Index (UQI) [Wang and Bovik, 2002]) comparing their results to the results of using these two methods, which indicate their method as being superior. However, these results are from a relatively small dataset ($n = 4$). This method was also used for the purpose of registration in [Chen et al., 2015b], and a similar approach used for the super-resolution of DTI images [Alexander et al., 2014]. Here, the regression forests learn a mapping from a $5 \times 5 \times 5$ low-resolution patch to a $2 \times 2 \times 2$ high-resolution patch for 6 diffusion tensor channels. Again, the authors of neither paper report a method of intensity normalisation.

This regression forest approach was developed further in [Jog et al., 2015]. The regression forest input was augmented with the addition of a descriptor designed to give spacial context to the patch. The outputs of these forests are used to form a pre-trained conditional random field which is used to infer the final synthesised image. As part of their paper, the authors show that this method can be used to fully synthesise pathological FLAIR images (in contrast to [Roy et al., 2013]) from T_1 images, and that segmenting lesions from these images with a standard tool yields more accurate results than attempting to segment them from the corresponding T_1 images. This approach was further developed in [Jog et al., 2017] by introducing a multi-resolution approach, where features are extracted and relationships learned at different resolution levels, building a synthetic image from course to fine features.

Other approaches

Deformable atlases were proposed as a method of generating a subject-specific image of a particular modality in [Miller et al., 1993]. The method involves using deformable registration to warp an atlas image from a target modality to the individual anatomy of the subject provided in a source image.

A model-based approach, whereby the intrinsic physical properties of the tissue being imaged are estimated from the available modalities was proposed in [Fischl et al., 2004]. The major problem with this and similar approaches is the requirement to have MR scans taken using particular sequences available. In fact, the authors also propose a novel acquisition sequence from which it is easier to extract the relevant information. Naturally, there are therefore very few images available acquired with this sequence, and the method can therefore not be used retrospectively. Whilst being very different from the more data-driven approaches described in this survey, it nevertheless provides an interesting alternative.

A method based upon registration and intensity fusion was proposed in [Burgos et al., 2014]. The method is used to synthesise CT from T_1 MR images for the purpose of attenuation correction for the reconstruction of PET/MR scans. The method can be briefly described as follows. First, a set of MR and CT images from the same subject are registered to each other using an affine transformation. Each MR image in this dataset is registered using a deformable registration to a source MR image for which a corresponding CT is to be synthesised. The Local Normalised Cross-correlation (LNCC) is subsequently calculated between the source image and each of the registered MR images. This measure provides a local similarity between each pair of images at each voxel. The CT images are then transformed using the same respective transformations and a spatially variant weighted average of each of them is computed using the LNCC maps computed previously as weights, yielding the final synthesised image.

Another approach to synthesis was proposed in [Vemulapalli et al., 2015], whereby synthesis is to be carried out in what the authors refer to as an "unsupervised" setting. The difference between this setting and those described in the papers surveyed so far in this section is that

there is assumed to be no training data where both the source and target modalities for the same subjects are available. Instead, only data from the target modality is available. The method proposed has two steps. In the first, a set of candidate intensities for a particular voxel are sought by finding the closest matching patches in the set of target modality images to a given patch in the source image. The similarity metric used is Mutual Information (MI) which is robust to comparisons between different modalities. The intensities of the centre voxels of these patches are thus taken as the candidate intensities for the respective voxel location in the synthesised image. The next step is to choose the best candidate voxels, whilst maximising spatial consistency. This is done by selecting from the candidate intensities at each voxel the weighted combination of intensities which will minimise a cost function consisting of a weighted sum of two terms. The first being the MI between the resulting synthesised image and the source image, the second being a metric measuring spatial consistency. This can be solved using gradient descent. The resulting image is then refined through coupled sparse representation.

The authors also apply their method to the standard setting where both source and target modality training images are available, leading to an algorithm which develops the basic IA framework from [Hertzmann et al., 2001]. They compare both this and their unsupervised approach to the methods proposed in [Ye et al., 2013] and [Van Nguyen et al., 2015], reporting comparable correlation and Signal to Noise Ratio (SNR) for the unsupervised method, and superior results for the supervised approach.

The last approach surveyed in this category involves attempting to create a generative model which describes the joint probability distribution of observing a pair of source and target modality images given a set of previously observed image pairs [Cardoso et al., 2015]. The advantage of having a full model such as this is that an image can be synthesised along with an uncertainty. Not only can the most likely target modality image for a given source be found, but each voxel can be associated with a distribution describing the confidence in the synthesised image. The authors use this method to synthesise a CT from a T_1 image, reporting results which compare favourably to those of [Burgos et al., 2014].

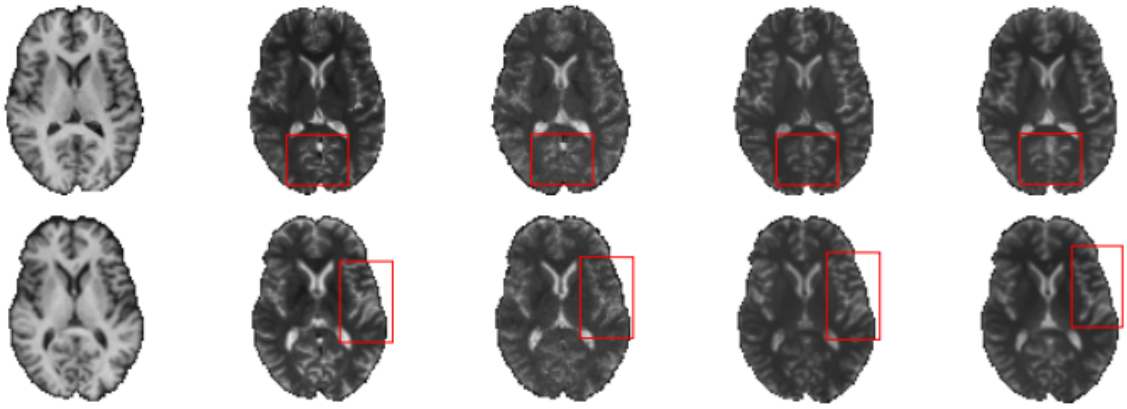


Figure 3.3: Two sets of images synthesised using three different methods, reproduced from [Van Nguyen et al., 2015] with permission. Left to right: Source T_1 image, real T_2 image, synthesised T_2 using method proposed in [Ye et al., 2013], synthesised T_2 using concatenated spatial information, synthesised T_2 using the proposed method. Red boxes indicate locations of significant difference.

Deep learning (non-GAN)

A deep learning approach was proposed in [Van Nguyen et al., 2015]. A location sensitive deep neural network was trained to synthesise T_2 from T_1 images and vice versa. The method follows a patch based approach in which target modality voxel intensities are synthesised from source modality patches independently of each other. The main deviation from standard deep learning architectures was the introduction of spatial information to the input of the network in the form of the voxel location. The spatial information is introduced into the network in a manner such as to tune individual sub-networks to particular spatial locations, which are subsequently switched on or off during synthesis. The authors showed that this approach yielded better results than simply concatenating the spatial information onto the intensity information to form the input layer. The paper includes an in-depth comparison to two other methods [Ye et al., 2013, Cao et al., 2014], showing their method to be both much faster and able to produce images with a significantly higher SNR. Figure 3.3 shows an example of synthesised T_2 images using a number of methods.

An alternative was proposed in [Sevetlidis et al., 2016] where modality transformation was performed using an autoencoder-like architecture. The first half of the network encodes the content of the source image to a latent representation, the second half reconstructs an image of the tar-

get modality from this latent representation. The intuition behind this approach is that a latent representation of the source image must contain all the relevant information about anatomy and structure. Reconstructing the target modality from this is then decoupled from the input modality. This was exploited in [Chartsias et al., 2018] where a multi-modal network for reconstruction was proposed. The authors encode multiple source modalities and fuse their latent encodings. A number of different target modalities can then be reconstructed from this latent representation. Fusion was performed using a pixel-wise maximum operator, meaning that missing input modalities could be handled while still producing a sensible fused encoding. In [Roy et al., 2017], head CT images were synthesised from ultra-short echo-time dual-echo MR images, again for the purpose of attenuation correction, using a CNN based on inception blocks [Szegedy et al., 2015].

Deep learning (GAN)

GANs have also been widely applied in the field of modality transformation. The methods described in Section 3.1.1 focused on pure image generation, however perhaps the greater impact of GANs has been the concept of an adversarial loss. This use of a separate network to effectively learn a loss function has been credited with much of the improvement in image quality seen compared to other generative models. This idea of an adversarial loss has therefore been applied, separate from the generative component, across many applications within the conditional image generation paradigm.

[Ledig et al., 2016] showed that an adversarially trained network could be used for highly realistic super-resolution with the SRGAN, while [Yoo et al., 2016] added two discriminators to provide both an image quality based loss and image relevance based loss for their domain transfer network. pix2pix [Isola et al., 2017] uses an adversarial loss combined with a traditional loss to perform supervised (with paired training images) style transfer. This was extended in CycleGAN [Zhu et al., 2017a] to use two adversarially trained networks for unsupervised (without paired training images) style transfer. By training both a forwards (style A to style B) and backwards (style B to style A) transfer network, they were able to intro-

duce an additional cycle-consistency loss reflecting the fact that if an image is transferred from style A to B and back again to A, it should be unchanged from the original image. By penalising networks which break this consistency, they were able to perform style transfer without the need for paired images. This approach has also made its way to modality transformation in medical imaging, with adversarial loss driven networks producing PET images conditioned on CT and ROI label maps [Bi et al., 2017], CT from MR images [Nie et al., 2016] and normal from low dose CT images [Wolterink et al., 2017]. Generating PET images from CT using a conditional GAN was investigated in [Ben-Cohen et al., 2017] and compared to results using a fully convolutional network, with the conclusion that the fusion of the two worked better than using the methods individually. Unsupervised synthesis using CycleGAN has also been investigated [Chartsias et al., 2017, Jin et al., 2018] for MR and CT synthesis. In [Huo et al., 2018], the authors propose a novel method for image segmentation without ground truth in the target modality. They propose using a CycleGAN to learn to transform from given modality A to a second modality B for which ground truth segmentations are available and perform segmentation on the synthetic image. They show that both the GAN and segmentation network can be trained simultaneously leading to improved segmentation results. Conditional GANs have also been used for MR super-resolution [Chen et al., 2018c, Sánchez and Vilaplana, 2018], registration [Dwarikanath et al., 2018, Yan et al., 2018] and the reconstruction of MR [Quan et al., 2017] and PET [Wang et al., 2018] images.

Driven by the success of the adversarial loss function in GANs, many authors have begun to include an adversarial component in segmentation tasks. The advantage of incorporating an adversarial loss into these methods is that higher-order features of true segmentations can be learned, such as topology, and the generated segmentation maps can be constrained to reflect these. The general architecture of these approaches involve using an image-to-image network such as a UNet [Ronneberger et al., 2015] to perform segmentation in a standard way, while also adding a discriminator to evaluate the realism of the generated images. The UNet (generator) is then updated using a combination of a typical cross-entropy segmentation loss on the ground truth images, as well as the adversarial loss from the discriminator network.

This approach has been extensively applied to data from the series of MICCAI BRATS work-

shops, which challenge participants to segment brain tumours on MR images into multiple classes. [Xue et al., 2017] uses the approach on data from the 2013 and 2015 editions, with [Rezaei et al., 2017] and [Li et al., 2017] applying similar approaches to data from the 2017 edition. The same method has also been applied to the segmentation of: MRI structures [Moeskops et al., 2017]; vessels in retinal fundoscopic images [Son et al., 2017]; optic disks and cups [Shankaranarayana et al., 2017]; organs on chest x-rays [Dai et al., 2017]; basal membranes [Wang et al., 2017]; skin lesions [Udrea and Mitra, 2017, Izadi et al., 2018]; prostate tissue [Kohl et al., 2017]; and the spleen [Huo et al., 2018] and liver [Yang et al., 2017]. All of these studies demonstrated improved segmentation performance by including the adversarial component.

Another approach to segmentation uses GANs as a means for unsupervised abnormality detection [Alex et al., 2017]. Here, the authors propose training a GAN on healthy image patches, and using the discriminator at the end of training to assign a likelihood to test patches as to whether it considers them real. Since the discriminator has only been trained on healthy images, it should assign patches from non-healthy regions a low probability of being real. These probabilities can then be used to identify regions of abnormality. When using such an approach, it is important to ensure sufficient training data is available and that the discriminator has sufficient capacity to learn the appearance of all non-pathological tissue. This can be a challenge as there is often a large amount of natural variation which must be encoded within the discriminator. Another interesting approach to segmentation was proposed in [Joyce et al., 2018], where segmentation is treated as a modality transfer problem from “MR image” to “segmentation map”. This allows for segmentation maps derived from an alternative modality to be used as the target modality, requiring no manual labels for the source modality images. Consistency between the generated segmentation maps and the source image must, however, be enforced, this is done by adding additional costs: an autoencoder-like loss to ensure a similar image to the original can be reconstructed from the segmentation maps, a term encouraging large segmentation maps, and a term minimising the variance of regions covered by the segmentation maps.

Discussion

Evaluation of synthesis methods: The question of how best to evaluate synthesis methods is difficult. Most synthesis methods have been employed as a preprocessing step prior to further analysis. As a result, the impact of the synthesis method can best be measured by comparing results of the final algorithm with and without synthesis, or to other state of the art methods for the individual applications. For example, methods which accurately synthesise high-frequency elements may yield results which are better for segmentation, whereas those which tend to accurately synthesise low-frequency components may be better for registration. Therefore the question of what's the best synthesis method may well be application specific.

Among the papers which aim solely for accurate synthesis, a variety of metrics have been used: MSE [Jog et al., 2013, Cardoso et al., 2015], correlation coefficient [Vemulapalli et al., 2015], SNR [Van Nguyen et al., 2015], UQI [Wang and Bovik, 2002] [Jog et al., 2013, Jog et al., 2015], peak SNR [Jog et al., 2015] and structural similarity [Wang et al., 2004, Jog et al., 2015]. Such a wide variety of metrics reported for a number of different applications and datasets makes it difficult to establish the state of the art, while the papers which do compare objectively against other methods papers [Van Nguyen et al., 2015, Cardoso et al., 2015, Vemulapalli et al., 2015], are not rigorous enough to establish a hierarchy of methods.

In order to establish a state of the art and have some means to compare new methods to it, it would be desirable to have a publicly available dataset to be used to establish a benchmark. Should image synthesis continue to be an area of development within the medical imaging community, it may become useful for there to be a framework within which different methods can be compared to each other. This may lend itself well to a challenge organised at a conference or workshop.

The role of registration: Accurate registration, whilst a long-standing problem which has received a huge amount of attention, is still a non-trivial task, particularly in the case of multi-modal data. In fact, several of the methods surveyed have used more accurate registration as a motivation for image synthesis, turning a multi-modal problem into a mono-modal problem.

Despite this, almost all methods involve registration of one form or another. The most common application of registration is in establishing a training dataset or patch library. In the latter, small misalignments may have significant consequences, yet these are rarely discussed in the literature. Having a fast and accurate method to find the closest matching patch in A for a given patch from B is not going to help if the patch then taken from A' is several voxels separate from the anatomy present in the patch taken from A.

A similar argument can be made for the methods which rely on regression. If there is misalignment within the training data, it is easy to foresee a case where the model will learn incorrect relationships as it relates intensities in one anatomical location to those in another. When developing a method which relies on registration, particularly multi-modal, it is important to consider what impact small registration errors will have and ensure the method is robust to them.

While most methods involve some form of registration, some approaches, particularly those which perform unsupervised synthesis such as [Vemulapalli et al., 2015] and those based on the CycleGAN approach [Zhu et al., 2017b], avoid the need for accurate registration through their use of unpaired image data, through a supervised approach tends to perform better where paired data is available [Vemulapalli et al., 2015].

Search speed solutions for patch matching methods: The most simple solution to the problem of search speed is to limit the size of library A to a single [Hertzmann et al., 2001, Roy et al., 2014b, Roy et al., 2011] or pair [Roy et al., 2010] of atlases. The drawback of this is that there may not be enough variability in A to capture all of the cases which can occur in B. This may be acceptable when synthesising non-pathological images, where there is likely to be little inter-subject variability at a patch level, however, it will be of a particular problem in cases where B displays some pathology which is not captured in A.

Whilst potentially reducing the search space significantly, restricting the search by using tissue segmentation, as described in [Roy et al., 2011], will cause the results to become dependent on the quality of the segmentation. When it comes to synthesising patches which have been wrongly segmented, the search for similar patches among those of a different tissue type will

yield poor results. When the problem to be solved is that of improving segmentation accuracy, as is the case in [Roy et al., 2011], the problem becomes circular. A time consuming iterative approach may, therefore, be required to reach the true potential of the algorithm.

Co-registering prior to a local search, as used in [Iglesias et al., 2013, Konukoglu et al., 2013, Ye et al., 2013, Tsunoda et al., 2014], will ensure that the search will be carried out among a selection of representative candidate patches. However, because the appearance of WM and GM is fairly homogeneous across an image, it is unlikely that the globally most similar patch will lie in the local area being searched.

Considering these latter two approaches, it is possible to come to the conclusion that the former will have a worse worst-case situation, in the rare case where you attempt to synthesise from an incorrectly segmented patch, whilst being close to globally optimum in the other cases. The latter may not find the globally optimal solution in the majority of cases but will be robust enough to handle all cases reasonably well. The choice between the two will, therefore, depend on the application, and whether it is better to have a globally more optimal solution, with occasional artefacts, or an artefact free, but less optimal solution. Prevalence in the literature suggests that the latter approach is more popular.

Random sampling, as used in [Rueda et al., 2013] should be seen as a last resort for the purposes of limiting the size of a search when there is no prior information available which could be used to otherwise guide the search. Even then, sampling from a more carefully selected subset may yield better results. Such a subset could be chosen based upon having a similar mean and/or variance to the given patch, for example.

Choice of similarity metric for patch-based methods: The majority of the patch based approaches to synthesis require the choice of some sort of similarity metric. Most of the methods surveyed here have used the L_2-norm, however, some have chosen to use a Normalised Cross-correlation (NCC). The authors in [Tsunoda et al., 2014] justify their choice of NCC by pointing to its relative invariance to linear intensity differences, which is necessary for their application to chest radiographs where bulk changes in intensity across an image is common. Bias field corrected brain images tend not to have these problems, and so many methods with applications

to brain images choose to use the L_2-norm. There may, however, be occasions when NCC could be preferable. Bias field correction is an important preprocessing step in many algorithms which aim to remove the low-frequency changes in intensity associated with magnetic field inhomogeneity within the MR scanner. However, bias field correction algorithms when applied to FLAIR images which contain large volumes of hyperintense lesions can lead to a reduction in contrast between the lesions and the surrounding tissue. It may, therefore, be desirable to perform synthesis using images which have not been bias field corrected, and thus exhibit bulk changes in intensity across the image. In this case, NCC may prove to be a more sensible choice of similarity metric.

Pseudo-healthy image synthesis: The idea of pseudo-healthy image synthesis forms the foundation of [Ye et al., 2013] and [Tsunoda et al., 2014], whilst being touched on as an unintended consequence of FLAIR synthesis in [Roy et al., 2013]. This is an interesting topic as the ability to synthesise images of healthy appearance from pathological images has potential applications not only in abnormality detection but also in other areas such as registration. The FLAIR images synthesised by the MIMECS algorithm in [Roy et al., 2013] are not truly pseudo-healthy as they do still provide contrast between lesions and the surrounding WM. The explanation provided is that, because the lesions have a similar intensity to GM in T_1 images, they are synthesised using patches taken from GM, and therefore have the slightly hyperintense appearance of GM in the synthesised FLAIR image. This is partly a consequence of the global search employed by MIMECS being able to match patches from anywhere in the image. It may also be a consequence of the segmentation approach, with inaccurate segmentations of lesions, a common occurrence in segmentation algorithms, leading to patches being searched for in the wrong library.

Intensity normalisation: An important problem which many papers fail to propose a solution to is that of intensity normalisation. In order to be able to accurately relate intensities between images, it is important to ensure that a given intensity means the same in each image. This is less important in some cases, for example when using CT images as CT data is fully quantitative. It is also less important when alternative similarity measures to the L_2-norm are used. For regression methods, however, this is a vitally important step, the details of which are omitted

in many such papers surveyed here.

Whilst MR images are not strictly quantitative, they are measures of physical parameters. Two images of the same tissue should, therefore, be approximately linearly related, assuming the impact of any bias field is minimal or can be removed prior to normalisation. It can, therefore, be assumed that the optimum function to map intensities of different images to a standard space will be approximately linear. The normalisation method used in [Rueda et al., 2013] of scaling each image to have intensities between 0 and 1 violates this under all but the most ideal circumstances, as it assumes identical intensity distributions between images. Any inter-image variation in the maximum or minimum intensities relative to the key features of the intensity histogram (for example, peaks corresponding to tissue types) will result in these features being mapped to different intensities in different images. While the minimum intensity in an MR image is fairly robust, the maximum is not and can be strongly affected by noise, artefacts or pathology.

Matching the mean and standard deviations of each image, as used in [Hertzmann et al., 2001], is more robust to noise, but will be vulnerable to variations in anatomy. The relative distribution of tissue types in the brain is subject specific and is dependent on factors such as age and pathology. This distribution will affect both the mean and the standard deviation of the intensities in an image.

Finding the peak of the histogram of intensities in the WM and setting this to 1 as used in [Roy et al., 2011, Roy et al., 2013] preserves the suspected linear relationship and is robust to noise. However, pathology present in the WM can affect the peak value used, especially in extreme cases.

Histogram matching, proposed in [Ye et al., 2013] should, under ideal circumstances, result in a linear transfer function. However, like matching the peak of the WM histograms, this approach will be strongly affected by pathology, particularly in the case of white matter hyper-intensities visible in FLAIR images.

Another method, used in [Huppertz et al., 2011] where the aim is to perform normalisation

in the presence of pathology, looks to define a robust fixed point which can be calculated for all images and scaled to an arbitrary value. The authors use a probabilistic segmentation to estimate the locations of WM and GM, after which the set of voxels for which the probabilities of being WM or GM are high are selected. Within these two sets, outliers are removed by treating the intensity distributions as Gaussian and keeping only those intensities which lie within a 95% confidence interval. The means of these two sets are then calculated, providing a robust mean of the WM and GM intensities. The mean of these two values is then calculated, and a scaling is applied to the whole image such that this value will be set to an arbitrary fixed number.

This approach aims to be robust to segmentation errors, common in pathological cases, by only using the values for which there is a high degree of confidence that they are from WM or GM. It then looks to be robust to intensity modifying pathology within these tissues by only using values which fall within a confidence interval.

Possible areas of future investigation: A common technique when dealing with patch libraries is to augment them with reflections or rotations of the original patches, or to search them using a rotationally invariant similarity measure [Grewenig et al., 2011]. None of the methods reviewed here do that. It is possible that this is unnecessary due to the nature of the images being synthesised. CSF, WM and deep GM tend to be very homogeneous, and therefore inherently rotationally invariant at the patch sizes typically used. Cortical GM is less homogeneous, however, its folded structure means that there will be patches present at all angles, meaning the library will already contain such patches without the need for augmentation. This is true also for the boundaries between GM and WM. As these cases cover the majority of MR images there may be little benefit to considering rotations or reflections within the patches, however, the subject could be investigated, as it could allow for larger patch sizes to be used.

As detailed above, neither limiting the patch search space though segmentations or through locality is ideal. A probabilistic segmentation may help address the problems associated with the former, with samples taken from the sets of different tissue types according to the probability of the given patch being of that type. A combination of the two methods may also prove useful,

with local search used to ensure a reasonable match is found, and tissue-based search increasing the chance of the optimal match being found.

Whilst not strictly image synthesis in the way presented here, a similar problem is presented in [Zhang et al., 2012] where the aim is the super-resolution of 4-dimensional (4D) CT lung images, consisting of a temporally varying 3D volume. The method proposed involves searching a library of patches for the most appropriate ones to populate a slice between acquired slices. Instead of using a static library, the method dynamically creates bespoke small libraries based upon an estimation of where the anatomy looking to be resolved was likely to be at other time points. Whilst not directly applicable in most applications described here, this solution could provide inspiration for work in other areas. For example in longitudinal studies of atrophy, where an approximately true copy of the modality to be synthesised may be available at a different time point. This approximately true image could then form a bespoke library from which to synthesise the modality at other time points. One possible application for this could be "updating" old scans of a patient in a longitudinal study when a change in protocol leads to a different scanner or imaging sequence being used, such is the case in the BLSA study [Shock et al., 1984]. As demonstrated in [Roy et al., 2011, Roy et al., 2013], segmentation algorithms are often more accurate on images acquired with particular protocols. Another application could, therefore, be used to condition the older scans for use with a particular algorithm. In both cases, the properties of a single image acquired with the new protocol could be propagated back to the older scans.

3.2 Lesion Segmentation

WMH are commonly found in FLAIR MR images. Their aetiology is diverse but they are known to be associated with a greater risk of stroke, dementia and death [Debette and Markus, 2010]. The quantification of WMH for use as a biomarker for diseases such as multiple sclerosis (MS) [Fu et al., 1998], dementia [Barber et al., 1999] and diabetes [Tamura and Araki, 2015] has lead to a large number of techniques for automatic WMH segmentation to be developed.

A consequence of their diverse aetiology and the desire of automated segmentation approaches to remove the need for manual segmentations in large scale studies and diagnostics is that there exists an extremely large body of literature on the subject. A 2013 survey of lesion segmentation methods applied to MS lesions alone found 80 papers which proposed an automated method, of which 47 included quantitative analysis [Garcia-Lorenzo et al., 2013]. A comprehensive review of these methods, along with those proposed since 2013 and those proposed for applications in other fields would be extremely large and beyond the scope of this theses. The methods described here are therefore limited to those which perform FLAIR image segmentation, and for which implementations are readily available, and as such, provide a clear source of comparison for any newly proposed method. First, however, we discuss how best to evaluate a novel segmentation method so as to provide the most useful information.

3.2.1 How to evaluate a segmentation method

One of the key issues raised in [Garcia-Lorenzo et al., 2013] is that each paper proposing a novel segmentation method will carry out an evaluation in a different way, with the main source of difference being the data used. The performance of most methods will depend heavily on the images used for evaluation. Each dataset will have images taken from different scanners at different field strengths, with different acquisition parameters and different subjects of varying age and degree of pathology. As such, any performance metrics pertaining to one method's application to a particular dataset cannot be compared directly to the metrics obtained when using another method on a different dataset. The most useful results come from studies which use multi-centre datasets for evaluation, demonstrating the proposed method's ability to generalise to heterogeneous data. However, only 13 papers in [Garcia-Lorenzo et al., 2013] reported this.

In order to directly compare methods, they must be compared on the same dataset. This provides a significant barrier to evaluation as the majority of authors do not make the code for their method publicly available. The authors of a new method will likely not have the time or desire to implement a large number of alternative methods. Even when the code is made

available, it can be difficult to install, require modification to be applied to particular datasets, or require the images to go through a time-consuming set of preprocessing steps.

The 2008 MICCAI MS lesion segmentation challenge [Styner et al., 2008] provided a platform for methods to be compared against each other using a standard dataset. This allowed for hitherto impossible comparisons to be made between a large number of methods on the same dataset, allowing for a clear hierarchy of methods to be formed. However, a drawback of a challenge such as this is that it provides no evidence of how a method will perform on the heterogeneous datasets seen in the real world. A method may be tuned such that it will perform very well on the cases and images provided by the challenge, yet be unable to generalise, whereas another method may not perform as well on the challenge dataset, but achieve the same level of performance on any other dataset. A future challenge could address this problem by including data from multiple sources, including real-world clinically acquired data from a variety of hospitals. However, it would be time-consuming and expensive to both acquire and label such a large dataset

Another barrier to inter-method comparison is the choice of the metric used to perform the evaluation. Many metrics have been proposed for the evaluation of segmentation a with corresponding surface S_a and volume V_a, with respect to a ground truth t with corresponding surface S_t and volume V_t, some of which are given below:

- *Dice Similarity Coefficient (DSC)*

  A measure of overlap between the volume of the computed segmentations and the corresponding reference segmentations [Dice, 1945]. Provides an overall measure of the accuracy of the computed segmentation, but becomes more sensitive to errors for small lesions. A DSC of 0 indicates no overlap, while a DSC of 1 indicates a perfect overlap.

  Defined as $\frac{2|V_a \cap V_t|}{|V_a|+|V_t|}$.

- *Jaccard Similarity (JS)*

Similar to DSC. A measure which is more sensitive to larger errors than DSC. Otherwise monotonically equivalent in that $J = DSC/(2 - DSC)$. Defined as $\frac{|V_a \cap V_t|}{|V_a \cup V_t|}$

- *Volume Similarity*

  A measure of the relative difference in total segmented volume between the computed and reference segmentation. Compares the size of the two segmentations with no indication of overlap. $\frac{2|V_a|-|V_t|}{|V_a|+|V_t|}$

- *Average Symmetric Surface Distance (ASSD) (mm)*

  A measure of the average distances between the surface of the computed segmentations and reference segmentations, and vice-versa. Provides an indication of how well the boundaries of the two segmentations align.

  Defined as $\frac{1}{2}(\Sigma_{t=S_t}(\text{mindist}(t, S_a)/|S_t|) + \Sigma_{a=S_a}(\text{mindist}(a, S_t)/|S_a|))$

  where $\text{mindist}(p, S)$ is the smallest Euclidean distance between surface point p and any point on S.

- *Hausdorff Distance (HD) (mm)*

  A measure of the maximal distance between the surfaces of the computed and reference segmentations. More sensitive to segmentation errors occurring away from segmentation boundaries than ASSD.

  Defined as $\max\{\{\text{mindist}(a, S_t), a \in S_a\}, \{\text{mindist}(t, S_a), t \in S_t\}\}$, where $\text{mindist}(p, S)$ is the smallest Euclidean distance between point p and any point in S and $\max\{A\}$ returns the greatest value in set A.

- *Precision*

  The proportion of the computed segmentation which overlaps with the reference segmentation. Provides an indication of over-segmentation. Ranges between 0 and 1.

  Defined as $\frac{|V_a \cap V_t|}{|V_a|}$.

- *Recall*

The proportion of the reference segmentation which overlaps with the computed segmentation. Provides an indication of under-segmentation. Ranges between 0 and 1.

Defined as $\frac{|V_a \cap V_t|}{|V_t|}$.

The following groupwise correlations can also be computed:

- *Intra Class Correlation (ICC)*

  A measure of correlation between $|V_t|$ and $|V_a|$. Calculated as ICC(A,1) defined as in [McGraw, K., Wong, 1996].

- *Scatter and Bland-Altman plots*

  Scatter and Bland-Altman plots showing the relationship between $|V_a|/|V_{ic}|$ and $|V_t|/|V_{ic}|$. The distribution of lesion volumes are often non-normal and hence non-parametric metrics are used. Scatter plots show how closely the two sets of values are related, with a low variance distribution along the line $y = x$ indicating a strong correspondence. Bland-Altman plots give a further measure of the agreement between the two sets of values, robust to sample selection [Bland and Altman, 2010]. $\frac{|V_a|/|V_{ic}|-|V_t|/|V_{ic}|}{0.5(|V_a|/|V_{ic}|+|V_t|/|V_{ic}|)}$ are plotted, with the desire for the mean to be close to zero, indicating a lack of fixed bias, and variance to be small, indicating a high degree of agreement. Visually it is also desired that there are no general upward or downward trends in the data which would indicate a volume dependent bias. These plots are associated with a number of metrics:

  - Equation of best fit line: Of the form $y = mx + c$, found by minimising the Sum of Squared Errors (SSE). Indicates how close the relationship between the two datasets is to the ideal ($y = 1x+0$). A larger value of $|c|$ indicates a constant error independent of lesion volume, while the a value of m differing from 1, indicates an error dependent on lesion volume.
  - SSE: Indicates how well the above equation fits the data.

  - r^2: The square of the Pearson correlation coefficient. Indicates how strongly correlated the two volume measures are with a value of 1 indicating a perfect correlation.
  - Reproducibility Coefficient (RPC): Indicates how well the automated method reproduces the results of the reference volumes.
  - Coefficient of Variation (CV): Indicates the strength of agreement between the two volume measures.
  - Mean: Indicates a fixed bias if different from zero. P-values signalling this difference are also usually calculated.

When proposing a new method, there are therefore a number of evaluation steps which should be taken to provide sufficient evidence of the method's performance.

- Apply a selection of publicly available methods to act as reference points.
- Evaluate the new method on data from multiple centres where possible to demonstrate its robustness to the heterogeneity found in real-world data.
- In the case non- disease specific methods, demonstrate their performance on a variety of datasets containing images of patients with different diseases.
- Perform evaluation using a range of metrics, and provide analysis with reference to the implications and limitations of each. Where possible, use direct (eg. DSC) and indirect (eg. ability for segmentation volumes to predict disease state) metrics to provide a measure of real-world applicability.

3.2.2 Publicly available segmentation methods

White matter hyperintensity segmentation

There are a number of publicly available methods for automatic WMH segmentation. This section contains a description of each and where they can be found.

The first pair of methods come from the Lesion Segmentation Toolbox[2] - the Lesion Growth Algorithm (LST-LGA) [Schmidt et al., 2012] and the Lesion Prediction Algorithm (LST-LPA). In the former, which requires both a T_1 and a FLAIR image, white matter, grey matter and CSF segmentations are first obtained from the T_1 image. These tissue segmentations are then used to create a lesion belief map from the FLAIR image. This is first thresholded by a value κ and the resulting segmentations are grown along hyperintense voxels. LST-LPA is a supervised method for which a logistic regression model was trained on 53 MS patients with severe white matter lesion loads. Both methods output a lesion probability map which the documentation suggests should be thresholded at 0.5.

The LesionTOADs [Shiee et al., 2010] tool, [3] as a plug-in for MIPAV [McAuliffe et al., 2001], combines the task of lesion segmentation with that of tissue segmentation by using an atlas-based approach. An initial segmentation estimate is acquired by registering the image with an atlas formed from the manual segmentations of a number of images and propagating the labels. This segmentation is refined through a combination of assigning intensity centroids for each tissue type and the successive shrinking and growing of the segmentation boundaries in order to minimise an energy function. A key feature of this method is that it maintains the topology of the segmentations from the initial atlas. Lesions will alter this topology and are therefore modelled as part of the WM for the purposes of topological refinement, but maintain their own intensity centroid.

A method[4] [Souplet et al., 2008] developed for the 2008 MICCAI lesion segmentation challenge, first performs tissue segmentation by using the EM algorithm with 10 classes - WM, GM, CSF, 6 GM/CSF partial volumes, and outliers mainly corresponding to vessels. Lesions are then found by first selecting the voxels in the FLAIR image which have an intensity greater than a threshold T such that $T = \mu_{GM} + 2\sigma_{GM}$ where μ_{GM} and σ_{GM} are respectively the mean and standard deviation of the previously found GM class. These voxels are used as seed points for subsequent expansion using morphological operations.

[2]available at: `www.statistical-modelling.de/lst`
[3]available at: `www.nitrc.org/projects/toads-cruise/`
[4]available at: `www-sop.inria.fr/asclepios/software/SepINRIA/`

Finally, there are three further methods [5]. These are STREM [Garcia-Lorenzo et al., 2008], MS4MS [García-Lorenzo et al., 2008] and GCEM [García-Lorenzo et al., 2009]. STREM uses a modified variant of the EM approach used above which removes the need for additional tissue classes beyond WM, GM and CSF by allowing a certain number of voxels to fall out of the model and be considered outliers. This encourages only those voxels which fit into the assumed 3 class Gaussian mixture to be classified, which leads to lesions being considered among the outliers. These outliers are refined by ensuring they are indeed sufficiently far from the 3 class distributions. Next, heuristic rules are imposed to remove false positives. First, the lesions are located among the outliers by selecting those with an intensity greater than $\mu_{WM} + 3\sigma_{WM}$ where μ_{WM} and σ_{WM} are respectively the mean and standard deviation of the WM class. Next, small ($< 3\text{mm}^3$) groups of connected voxels among the remaining outliers are removed. Finally, connected voxels which do not share a border with the WM, or share a border with the brain itself, are removed. The method was evaluated as part of the 2008 MICCAI lesion segmentation challenge. Notable among the results is the extremely high specificity (0.9954) at a cost of very low sensitivity (0.2562). This suggests that many lesions are being missed, perhaps as a result of the heuristic rules being too strict.

The introduction of a mean-shift [Fukunaga and Hostetler, 1975] process to perform an initial segmentation improved upon this method resulting in the MS4MS algorithm. Mean-shift is a method of finding a local maximum in an image. A segmentation can, therefore, be found by grouping voxels such that all voxels in a group share the same local maximum. In other words, from any point in this set, repeatedly travelling in the direction of greatest increase in intensity will lead to the same point. These regions are refined by merging nearby regions, with distance defined as the Euclidean distance between the corresponding maxima, and by merging small regions ($< 3\text{mm}^3$) with neighbouring ones. These regions are next classified as inliers or outliers by comparing their maxima to a Gaussian mixture model calculated as in STREM. Should a given maximum be found to have a sufficiently small p-value to belong to the model, then it is treated as a potential lesion. Finally, the same heuristic rules as used in STREM are used to limit false positives.

[5] available at: `www.irisa.fr/visages/benchmarks`

Finally, GCEM uses the results of STREM as initialisation of a graph cut procedure, whereby a segmentation is found by applying a max flow algorithm [Boykov and Funka-Lea, 2006] to an undirected graph where each node represents a voxel which is connected to its neighbours and two special source/sink nodes. The objective is to find a labelling such that an energy function is minimised. This function is a weighted combination of a regional term, which ensures voxels having the same label have a similar appearance, and a boundary term, which ensures that no boundaries exist between a set of voxels with the same label. The results of STREM allow for an initialisation whereby voxels in the estimated lesions are connected to the sink node, and the other voxels are connected to the source node.

Non-lesion segmentation methods

There are many segmentation tools which are publicly available for other types of segmentation tasks. Work in this thesis makes extensive use of two of these: MALPEM [Ledig et al., 2015] and DeepMedic [Kamnitsas et al., 2017b]. MALPEM is a structural segmentation tool which segments a T_1-weighted brain image into 139 anatomical structures. It is designed to be robust to pathology and has therefore been used in studies of traumatic brain injury [Ledig et al., 2015] and dementia [Ledig et al., 2018b]. MALPEM works by co-registering a number of labelled atlases onto the target anatomy and propagating the labels from the atlas to the target image. Labels from across the atlases are fused to give a single label for each voxel, which is then refined by intensity using the EM algorithm. Such segmentations are valuable in many processing pipelines, including intensity normalisation using WM and GM masks and the construction of anatomical regions of interest. The publicly available implementation of MALPEM also encapsulates a number of useful preprocessing steps: intensity inhomogeneity correction using the N4 algorithm [Tustison et al., 2010], brain extraction using the PINCRAM algorithm [Heckemann et al., 2015] and affine registration to a standard Montreal Neurological Institute (MNI) space. Preprocessing using the MALPEM tool is performed on all T_1-weighted MR images used in this thesis, yielding bias corrected, brain extracted, co-registered images, with accurate structural segmentations.

DeepMedic is a general purpose segmentation network architecture which has been shown to provide accurate multi-class segmentation results out-of-the-box with minimal modifications across a number of tasks. It achieved the highest results in the 2015 ISLES ischemic stroke lesion segmentation challenge, and formed part of an ensemble of methods [Kamnitsas et al., 2017a] which won the 2017 BRATS challenge for brain tumour segmentation. It's also achieved competitive results in WMH segmentation [Guerrero et al., 2018]. The network is primarily configured for 3D segmentation, though modifications required for 2D are presented in [Kamnitsas et al., 2017b]. Chapters 5 and 6 of this thesis employ DeepMedic as a reliable segmentation network architecture to evaluate the effects of different approaches to data augmentation.

3.3 Data augmentation

Data augmentation is the process of expanding a training dataset by including additional data derived from the available real data. In the case of imaging data, common methods for data augmentation include rotation, translation, scaling, intensity scaling, reflection and random deformations. The aim of data augmentation is to increase the number of training images by introducing additional feasible data points. In the case of rotation augmentation, the real images are rotated through a set of random angles, with the generated images introduced into the dataset. Translation augmentation, scaling and random deformations are similar, novel images are generated by applying random translations, scaling factors, or elastic deformations to the original images. When using reflection augmentation, the available images are simply reflected along one or more predefined axes. Finally, intensity augmentation involves generating new images by changing the pixel intensities of the originals. This can be a simple scaling or can aim to make more complex changes such as altering lighting conditions. The main benefit of all of these methods is to reduce the potential for overfitting in a model trained on the data. For example, in a small dataset, it is highly likely that irrelevant information such as the angle of an image is coincidentally correlated with an image label. Learning this correlation will cause the learning algorithm to incorrectly classify images at a different angle. By artificially rotating the training images and including them in the training data, the coincidental correlation is removed

and the learning algorithm must identify other features upon which to base its classification.

Simple forms of data augmentation such as rotation and translation can be considered a basic form of image synthesis, where hitherto unseen cases are synthesised from the existing cases. Used correctly, these simple methods are relatively low-risk, being unlikely to reduce learning performance. Other forms such as random deformations are higher risk and their use should be considered carefully. For example, augmenting using random deformations could reduce performance in tasks where straight lines are a strong discriminative feature. Applying the right forms of augmentation can lead to significant increases in performance [Krizhevsky et al., 2012, Ronneberger et al., 2015]. Choosing the right method for data augmentation is, therefore, an important step when developing or applying a learning algorithm. In practice, the choice as to what forms of augmentation will be useful, and which may be potentially detrimental, comes from domain knowledge and experimentation. For example, rotation augmentation should not be applied when training a classifier where the orientation of image components was known to be a discriminative feature. If such information is not known a priori, a series of experiments could be performed to either ascertain whether orientation is indeed discriminative or to directly assess the impact of the augmentation regardless.

As noted in [Krivov et al., 2017], the consensus within the medical imaging community appears to be to forgo extensive data augmentation in favour of preprocessing, particularly in neuroimaging. The authors argue that the extensive range of preprocessing options available for neurological images removes the need for many forms of data augmentation - why apply rotation, translation and scaling augmentation when images can simply be co-registered, removing all that irrelevant variation? Intensity augmentation is also made redundant through bias correction and intensity normalisation. This is a fairly unique property of brain MR images, where, pending incremental improvements in reliability, normalisation is almost a "solved problem". There are a number of different tools available to address each of these sources of variation, reflecting different needs for the degree of normalisation required and influence of pathology. However, one exception is protocol normalisation, where the goal is to remove the effects of different imaging protocols (as in [Roy et al., 2011]), which remains an area of active research. Such steps are significantly harder to perform on natural images, where authors tend

to fall back to data augmentation. One departure from this is in facial recognition tasks, where facial alignment is a common preprocessing step, though differences in lighting conditions are often still present and require addressing. The most common form of augmentation in neuroimaging is, therefore, reflection along the brain midline, as this will almost always provide sensible synthetic images.

Random deformations have been shown to be useful in some biomedical image segmentation tasks [Ronneberger et al., 2015, Milletari et al., 2016], however it is worth noting that these applications (cell and prostate segmentation) involve relatively non-rigid structures, while the authors of [Krivov et al., 2017] argue that applying such augmentations in neuroimaging would likely counteract the extensive normalisation procedures and potentially introduce cases of unrealistic anatomy. Despite this, there are likely theoretical advantages to applying some form of deformation augmentation to neurological images, however realising these in practice would involve the careful definition of what deformations are valid, and a practical way to generate them. The authors of [Krivov et al., 2017] propose an alternative, which involves propagating lesions from pathological images onto a number of healthy images, with promising results.

Very recently, GANs have begun to be proposed as an alternative way to perform data augmentation, with Chapters 5 and 6 of this thesis investigating this in detail.

The sources of variance in a dataset can be thought of as the intrinsic dimensions of the data. Each image can be represented as a point in an n-dimensional space, with travel along each dimension corresponding to changes in a different mode of variation. Many dimensionality reduction methods exist to try and infer this underlying latent space, such as PCA [Wold et al., 1987], Laplacian eigenmaps [Belkin and Niyogi, 2003], and t-SNE [Maaten and Hinton, 2008], however these require an estimate of the dimensionality of this space, i.e. how many sources of variance there are, and do not always provide a method to back-project these points to their original space to sample novel data points. Estimating this dimensionality can be difficult, especially in highly non-linear domains such as images. While ways exist to this [Ceruti et al., 2012, Lombardi et al., 2011], these work best in low dimensions. GANs therefore have an advantage

over such methods, as they do not require a strict estimate of the dimensionality of the underlying distribution and will work provided $|\mathbf{z}|$ is sufficiently large.

One of the earliest cases of using GAN derived synthetic data to augment training data was proposed in [Chartsias et al., 2017]. The authors use a CycleGAN to learn to map between unpaired labelled CT and MR images, thereby allowing labelled CT images to be transformed into labelled MR images. This effectively allows for CT images to be used as MR training images, leading to a 15% increase in DSC. The quality of the generated images was also evaluated and was found to be only slightly less valuable than real images, leading to a reduction in DSC of only 5% when used in place of real images. A method like this can have many applications in medical imaging, where there are many small labelled datasets available from different modalities and different acquisition protocols. Being able to combine all such images of the same pathology together using a series of CycleGANs is an exciting prospect. A similar approach was presented in [Shrivastava et al., 2016] for the purposed of gaze estimation and hand pose estimation, where instead of a separate dataset, the authors use a set of simulated images. Like before, a conditional GAN is trained to apply a style transfer, in this case from simulated appearance to realistic appearance, in order for the simulated images to be included directly into a training dataset. The advantage of using simulated images is that, given a suitable model, they can be produced "to order" with specified characteristics, and therefore do not require manual annotation. Another similar approach is made in [Mok and Chung, 2018], effectively combining the previous two methods. A conditional GAN is used to map from a semantic segmentation map to an MR image in a tumour segmentation task. The authors discuss the issues regarding applying random deformations directly to MR images, and instead apply the deformations to the segmentation maps containing both pathology and brain segmentation masks. The GAN then converts this into an MR image in such a way as to ensure that the final image is anatomically sensible (on the assumption that the GAN does not know how to produce anything else). In this way, random deformations, effectively a model of variation, can be used without risking image realism. While it is important to ensure the generation of realistic pathology, it is also important that anatomic sensibility is enforced. For example, generated images much follow basic anatomical rules such as the relative position of, and cor-

relations between, particular structures. A well-trained discriminator should be able to detect these anatomical abnormalities and ensure the generator does not produce such images. This does, however, rely on training a sufficiently powerful discriminator, requiring a large enough network and adequate training data.

A slightly different approach was taken in [Zhu et al., 2018]. Here, the authors use a conditional GAN to learn to impose specific emotions upon neutral faces for data balancing in an emotion classification task. A conditional GAN is used to learn a mapping from a neutral expression to one of an underrepresented class. This is essentially a form of style transfer, where styles are defined as emotional expressions. In this way, additional images can be generated with the broad characteristics of the neutral faces but portraying different emotions. It is not difficult to see such a method having an application in medical imaging - a conditional GAN could be learned to add pathology to otherwise healthy images in order to introduce greater anatomical variance to a small pathological training set. A somewhat opposite approach was also used in [Antoniou et al., 2017], where instead to imposing the relevant characteristic (eg. emotion) onto a set of images with irrelevant variation (eg. hair, eye, skin colour etc), the irrelevant variation is learned and imposed on an image displaying the relevant characteristic.

The majority of methods discussed here incorporate some additional knowledge into the data generation process in the form of a model (bold) or dataset (italics): In [Chartsias et al., 2017], the authors use *labelled CT images* as a source from which to generate labelled MR images; in [Shrivastava et al., 2016], the authors use **model** generated simulated images as source images; in [Mok and Chung, 2018], the authors use **random deformations** applied to real segmentation maps and produce images conditioned on these; in [Zhu et al., 2018], the authors use *faces displaying neutral expressions* as source images to produce images with a particular expression; in [Antoniou et al., 2017], the authors use a *related dataset* to learn a set of realistic modes of variation which can be imposed on a training image. While these are therefore no longer strictly pure data-driven augmentation procedures (using training data only), they do make an important point - that incorporating information from another source can potentially yield significant improvements beyond what's possible with pure data-driven approaches. This style of data augmentation is investigated further in Chapter 6 of this thesis.

Pure data-driven GAN based augmentation methods, where no additional information is provided beyond that which is present in the training image data, have received little attention when compared to the conditional approaches described above, with only a few very recent papers found to be making use of such techniques in medical imaging. In [Amitai and Goldberger, 2018], synthetic liver lesions are generated using a DCGAN architecture and used to augment a dataset for the purpose of lesion classification. A similar approach is taken in [Moradi et al., 2018] where synthetic normal and abnormal chest radiographs are generated using two DCGAN-like architectures. Both of these papers report that using GANs for data augmentation leads to improvements in classification accuracy for their respective tasks. In [Salehinejad et al., 2017], synthetic chest X-rays are generated in order to balance an imbalanced 5 class dataset. A separate GAN was trained on the available data from each class, after which synthetic images were sampled so as to have an equal quantity of images from each class present in the training dataset of a classifier, leading to significant increases in classification accuracy. Data balancing with GANs has also been proposed in [Mariani et al., 2018].

These are some intriguing results which merit further investigation. If indeed it is the case that GANs can be used as a simple prepossessing step, there may be potential to improve the training of learning algorithms in many applications, with no additional data being required. It also asks the question - How does a GAN learn to produce information which a classical (non-GAN) CNN cannot extract from the same data? The answer is unclear and warrants further investigation, but could be a consequence of loss function. The use of an adversarial loss could be leading to better internal representations of the data within the network. Alternatively, an adversarial loss means that GANs may be less prone to overfitting than classical CNNs, with their outputs being a broader representation of the data than a CNN would typically learn. Elucidating exactly why this is could lead to better loss functions being designed for classical CNN training. While in [Amitai and Goldberger, 2018, Moradi et al., 2018], GANs are used to grant an improvement by augmenting the training data in classification tasks, there are also likely improvements to be made in segmentation tasks. This is investigated in Chapter 5 of this thesis.

3.4 Is synthesis worth it?

The work in this thesis focuses on the topic of image synthesis, however, it could be argued that synthesis would not be necessary if learning algorithms were sufficiently trained with proper regularisation since synthesis itself invents no new data. This asks two questions in the context of this thesis. First, what is the advantage of using image-to-image synthesis, rather than learning directly from the modalities which are available? And second, how can training methods to generate synthetic training data lead to improvements in imaging tasks if such images contain no information which could have otherwise been learned from the original dataset?

To address the first question, we can consider that synthesis does allow us to indirectly inject more information into the learning process in the form of the synthesis model. Such a model must have been trained on examples of both modalities, and therefore encodes more information than a model trained on a single modality. This could, for example, be in the form of a spatial prior learned from across both datasets. The information contained in images which are then passed through this synthesis model becomes augmented with the additional information encoded within the model. Another benefit of synthesis is that it allows domain knowledge to be combined with a data-driven approach. For example, it may be known that a particular synthesis operator, which could be as simple as a Gaussian filter, increases the signal-to-noise ratio for the features one wishes to detect. It has been shown [Maier et al., 2018] that hard-coding these "known operators" is better than allowing a network to learn them for itself. This suggests that, if it is known that a particular form of image synthesis can be beneficial in detecting a particular feature, it is better to do this explicitly rather let the model learn it. Taken to the extreme, one could argue that a reconstructed image contains no more information that the raw data provided by an imaging device, however, learning from an image synthesised from this data is usually preferable to learning from the raw data itself.

The second question, regarding how synthetic data produced from a dataset improves algorithm performance when it contains no additional information beyond that which the dataset already contained, is more difficult. Where no new information is present, one would not expect an al-

gorithm trained on the synthetic data to outperform one trained on the original data. However, as described earlier, this effect has been observed in multiple studies. If this improvement does not stem from additional information, it must come as a result of the algorithms themselves. For example, training a GAN prior to a segmentation network means we are using three networks to process the data, therefore it could be that the improvement seen is simply a result of the training of more parameters. However, if this were the only reason, one would expect that increasing the size of the segmentation network would have the same effect, which is not the case. Instead, it is our belief that the improvement comes as a result of incorporating a different loss function, namely, the adversarial loss within the GAN. The mechanism for this is not clear, though it could be due to the adversarial loss being "softer" than traditional loss functions. In abstract terms, an adversarial loss ensures that "A must be like B", whereas traditional loss functions try to ensure that "A must equal B". In this way the training data is seen more like a continuous distribution as opposed to a set of discrete points, thereby acting against overfitting, particularly when the training dataset is small.

It is for these reasons that we believe the study of image synthesis is extremely relevant within the medical imaging domain, and that its appropriate use can yield significant improvements in multiple areas from across the discipline.

Chapter 4

Brain Lesion Segmentation through Image Synthesis and Outlier Detection

4.1 Introduction

Cerebral Small Vessel Disease (SVD) can manifest in a number of ways, many of which result in hyperintense regions visible on T_2 Magnetic Resonance (MR) images. The accurate automatic segmentation of these lesions is a key step in the diagnosis and study of SVD has been the focus of many of the methods reviewed in Chapter 3. However, these approaches tend to be limited to certain types of pathology, as a consequence of either restricting the search to the white matter or by training on an individual pathology.

This chapter describes a general approach to White Matter Hyperintensity (WMH) segmentation using modality transformation and outlier detection which is able to detect abnormally hyperintense regions on Fluid-attenuated Inversion Recovery (FLAIR) regardless of the underlying pathology or location. This approach uses a combination of image synthesis, Gaussian mixture models and one class support vector machines, and needs only be trained on healthy tissue.

A novel modality transformation method using kernel regression to learn the expected relation-

ships between T_1 and FLAIR intensities at each location within the brain is first presented. This method is particularly suited to the task of "pseudo-healthy" image synthesis in the presence of T_1 visible pathology. Subtraction of the pseudo-healthy FLAIR image from the acquired FLAIR image then gives an indication of pathology. Gaussian Mixture Models (GMMs) are then used to locate regions of the FLAIR image which are abnormally bright. These two pieces of information are combined with an SVD atlas within a one class classification framework and the output is post-processed using a Conditional Random Field (CRF).

The described method is unsupervised in the sense that it does not require any manually segmented ground truth images to train on, and is, therefore, less prone to overfitting than supervised methods. It is also flexible enough to segment a wide range of abnormalities without needing to be trained on examples of different pathologies. It does, however, need to be trained on non-pathological tissue. This can either be from images of healthy subjects, or from the regions outside of manual segmentations of pathological images.

The remainder of this chapter is structured as follows. Section 4.2 contains a brief review of the most relevant areas from Chapters 2 and 3. Section 4.3 describes both the synthesis method and how the resulting images are used to form lesion segmentations. Next, Section 4.4 describes a number of experiments which were carried out to compare the described method to three established methods, while Section 4.5 contains the results of these experiments along with a discussion. The chapter is then concluded in Section 4.6 with some discussion on potential avenues for future research.

4.2 Background

Of the proposed methods to segment White Matter (WM) lesions, very few are publicly available. Of these, the most common comparator methods belong to the Lesion Segmentation Toolbox (LST). The LST contains two methods, the Lesion Growth Algorithm (LST-LGA) [Schmidt et al., 2012] and Lesion Prediction Algorithm (LST-LPA). Both methods were developed for the segmentation of Multiple Sclerosis (MS) lesions which, while caused by a dif-

ferent process, appear similar on MR images (see Section 2.4). Due to these similarities in the appearance of MS and WM lesions, MS [Garcia-Lorenzo et al., 2013, Lladó et al., 2012] and WM lesion segmentation algorithms are often used interchangeably. As such, both methods from the LST are commonly used as benchmarks when evaluating hyperintense lesion segmentation algorithms. Another publicly available method is LesionTOADS [Shiee et al., 2010], which simultaneously performs both tissue and lesion segmentation in an unsupervised manner. At the moment, LST-LPA is the closest the field has to a readily available and robust gold standard, having been shown to consistently offer good results across a number of datasets despite being primarily an MS lesion segmentation tool. However, because of this, LST-LPA has a number of limitations when it comes to detecting other sources of hyperintensity, particularly those which extend into the grey matter such as cortical infarcts, as it employs a WM mask to restrict its search.

The idea of pseudo-healthy image synthesis, where the aim is to synthesise a pathology free subject specific image in a target modality, has been explored previously and is discussed in detail in Section 3.1.3. Pseudo-healthy image synthesis has been used in several applications: in [Ye et al., 2013] to perform tumour segmentation, in [Tsunoda et al., 2014] to detect lung nodules on Computed Tomography (CT) images, and was suggested as a potential method for WM lesion segmentation in [Roy et al., 2013].

Pseudo-healthy image synthesis is most useful when pathology is not visible on one modality, but visible on another. By synthesising a pathology free version of the pathological modality, abnormalities can be identified through subtraction. This can be a challenge in SVD where pathology can be visible on both T_1 and FLAIR images (Figure 2.7). In fact, existing methods have been demonstrated to synthesise hyperintensities [Roy et al., 2013, Jog et al., 2017], and even exploit this [Jog et al., 2015] for the purpose of lesion segmentation in the absence of FLAIR. However, this chapter shows that careful design of the synthesis algorithm does allow for a pathology free FLAIR to be synthesised in the presence of T_1 visible pathology.

4.3 Method

4.3.1 Overview

The proposed method treats the problem of lesion segmentation as an outlier detection task. The first stage is to produce two likelihood maps:

$\mathbf{L^{SYN}}$, is formed by synthesising a healthy looking FLAIR image from a subject's T_1 image. Subtraction of this synthetic FLAIR image from the subject's true FLAIR image produces a difference image which represents the likelihood of a FLAIR voxel intensity to be abnormal, given the subject's T_1 image and an expected pre-determined relationship between healthy T_1 and FLAIR intensities. This value is low in the presence of healthy tissue and high in the presence of pathological tissue.

$\mathbf{L^{FLAIR}}$, represents the likelihood for a given FLAIR voxel to be abnormal given a pre-computed GMM of expected FLAIR intensities at that location.

These likelihood maps are then combined with a White Matter Hyperintensity of Presumed Vascular Origin (WMH_{pvo}) probability atlas within a one-class classification framework to provide a single likelihood map reflecting the degree of abnormality at each voxel. Finally, a CRF is applied, resulting in a binary segmentation.

$\mathbf{L^{SYN}}$, $\mathbf{L^{FLAIR}}$ and the one-class classifier used to combine them all require a training set of healthy subjects. $\mathbf{L^{SYN}}$ a requires both T_1 and FLAIR images, whilst $\mathbf{L^{FLAIR}}$ and the one-class classifier require FLAIR images. There is no requirement for the three training sets to include the same subjects, however, it is practical to use the same set of FLAIR images. The T_1 and FLAIR images in this dataset are therefore referred to as $\mathbf{T^{train}}$ and $\mathbf{F^{train}}$ respectively.

4.3.2 Preprocessing

Preprocessing is required to normalise the images to a standard set of properties, ensuring subsequent steps are robust to the heterogeneous image characteristics found both within and

between medical imaging datasets. These preprocessing steps also compute a number of segmentations and transformations which are required in subsequent steps. Preprocessing is identical for both the training set and the images to segment, which are referred to from here as the test set.

Registration

Registration is performed using the MIRTK suite of registration tools [1]. A rigid transformation from the T_1 to FLAIR image space is first computed. A Free Form Deformation (FFD) [Rueckert et al., 1999] transformation (Resolution levels: 40mm, 20mm, 10mm, 5mm; Image dissimilarity measure = Sum of Squared Differences (SSD); Bending energy weight = .1) is then computed between the T_1 image in FLAIR image space and an Montreal Neurological Institute (MNI) template [2]. The inverse transformation is also computed.

Bias correction, brain extraction and anatomical segmentation

A multi-atlas based anatomical segmentation tool called Multi-Atlas-Label Propagation with Expectation-Maximisation based refinement (MALPEM) [Ledig et al., 2015] (described in Section 3.2.2) is applied to the T_1 image providing both binary and probabilistic segmentations of 139 anatomical structures. As part of the segmentation process, MALPEM applies bias field correction using the N4 [Tustison et al., 2010] algorithm and brain extraction using the PINCRAM algorithm [Heckemann et al., 2015], outputting the resulting T_1 image and brain mask. WM and Grey Matter (GM) probability maps are computed from the probabilistic segmentations.

Bias correction is performed separately on the FLAIR image using the N4 algorithm and the T_1 brain mask is transformed to FLAIR image space, re-sampled using nearest-neighbour interpolation and used to crop the FLAIR image.

[1] Available at: `biomedia.doc.ic.ac.uk/software/mirtk/`

[2] ICBM 2009a Nonlinear Symmetric, available at: `www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009`

Intensity normalisation

Intensity normalisation is an especially important procedure since many subsequent steps involve direct comparisons between voxel intensities across images from different subjects. However, the nature of hyperintense lesions means that several commonly used normalisation methods are inadequate. The often used approach of linear scaling of intensities to the range $[0, 1]$ with a certain percentage of the lowest and highest intensities saturated at 0 and 1 respectively [Cao et al., 2014] will result in different intensity mappings dependent on the volume of hyperintense lesions compared to the percentage of voxels saturated. Histogram matching [Ye et al., 2013] suffers similar problems in the presence of hyperintensities. Scaling images to have a zero mean and unit variance [Hertzmann et al., 2001] is also inadequate as the degree of hyperintensity will bias both the mean and the variance of the image.

To make intensity normalisation invariant to the degree of hyperintensity and atrophy common in elderly subjects, the method used in [Huppertz et al., 2011] is employed. Two sets of voxels corresponding to WM and GM are produced by filtering probabilistic WM and GM masks to include only voxels with a $> 95\%$ probability of being of that tissue class. Next, these two sets are further refined by intensity to contain only intensities which fall within a 95% confidence interval so as to remove outliers. This leaves two sets which are highly likely to contain WM and GM, and which are not outliers within these groups, therefore corresponding only to healthy tissue. The mean of each set of intensities is calculated to give the expected intensity of healthy tissue in the WM and GM. The mean of these two values is subsequently calculated to provide a single fixed point. Finally, image intensities are scaled linearly such that this fixed point is set to the arbitrary value of 1000.

This method is applied to both the T_1 image and FLAIR image, using the probabilistic WM and GM masks derived from the previously computed anatomical segmentations. In the case of the FLAIR image these masks are transformed to FLAIR image space and re-sampled using linear interpolation.

4.3.3 Training

In order to produce $\mathbf{L}^{\mathbf{SYN}}$ and $\mathbf{L}^{\mathbf{FLAIR}}$, two sets of models are trained. The first is a synthesis model that learns the relationship between T_1 and FLAIR intensities. The second is a GMM which learns the expected intensity distributions within a FLAIR image.

To account for imperfect tissue segmentation, common in the presence of hyperintense lesions, and for intensity variations within a tissue type, both sets of models are computed in a voxel-wise manner within MNI space. A separate model is produced for each voxel, computed using information taken from a patch around that voxel in each co-registered training image. The process of training both models is summarised in Figure 4.1.

Synthesis model

The key step for the computation of $\mathbf{L}^{\mathbf{SYN}}$ is the calculation of a pseudo-healthy FLAIR image from a subject's T_1 image. The proposed method uses voxel-wise kernel regression to learn a direct mapping between healthy T_1 and FLAIR intensities at each voxel.

A set of n training image pairs $\mathbf{T}^{\mathbf{train}}$ and $\mathbf{F}^{\mathbf{train}}$ are transformed to MNI space using the transformations calculated during preprocessing and re-sampled onto a 1mm isotropic voxel lattice. Intensities in $\mathbf{T}^{\mathbf{train}}$ are capped at a value t_{max}. At each voxel $\mathbf{x}$, two one-dimensional vectors $\mathbf{t_x}$ and $\mathbf{f_x}$ are formed from $\mathbf{T}^{\mathbf{train}}$ and $\mathbf{F}^{\mathbf{train}}$ respectively containing the voxel intensities from an a-by-a-by-a patch around $\mathbf{x}$ in each image, with each vector being of length na^3. A kernel regression model $\mathbf{M_x}$ with bandwidth h is computed relating $\mathbf{t_x}$ to $\mathbf{f_x}$ and evaluated at m equally spaced values k between 0 and t_{max}.

$$\mathbf{M_x}(k) = \frac{\sum_i^{na^3}(K((\mathbf{t_x}(i))/h)\mathbf{f_x}(i))}{\sum_i^{na^3} K((k-\mathbf{t_x}(i))/h)},\ K(p) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}p^2}. \tag{4.1}$$

Higher values of m and t_{max} result in more accurate synthesis at the cost of model size and computation time, whilst the number of voxels (na^3) must be sufficiently large to contain enough information to fit the model. Preliminary experiments showed that $m = 100, t_{max} = 1500, n =$

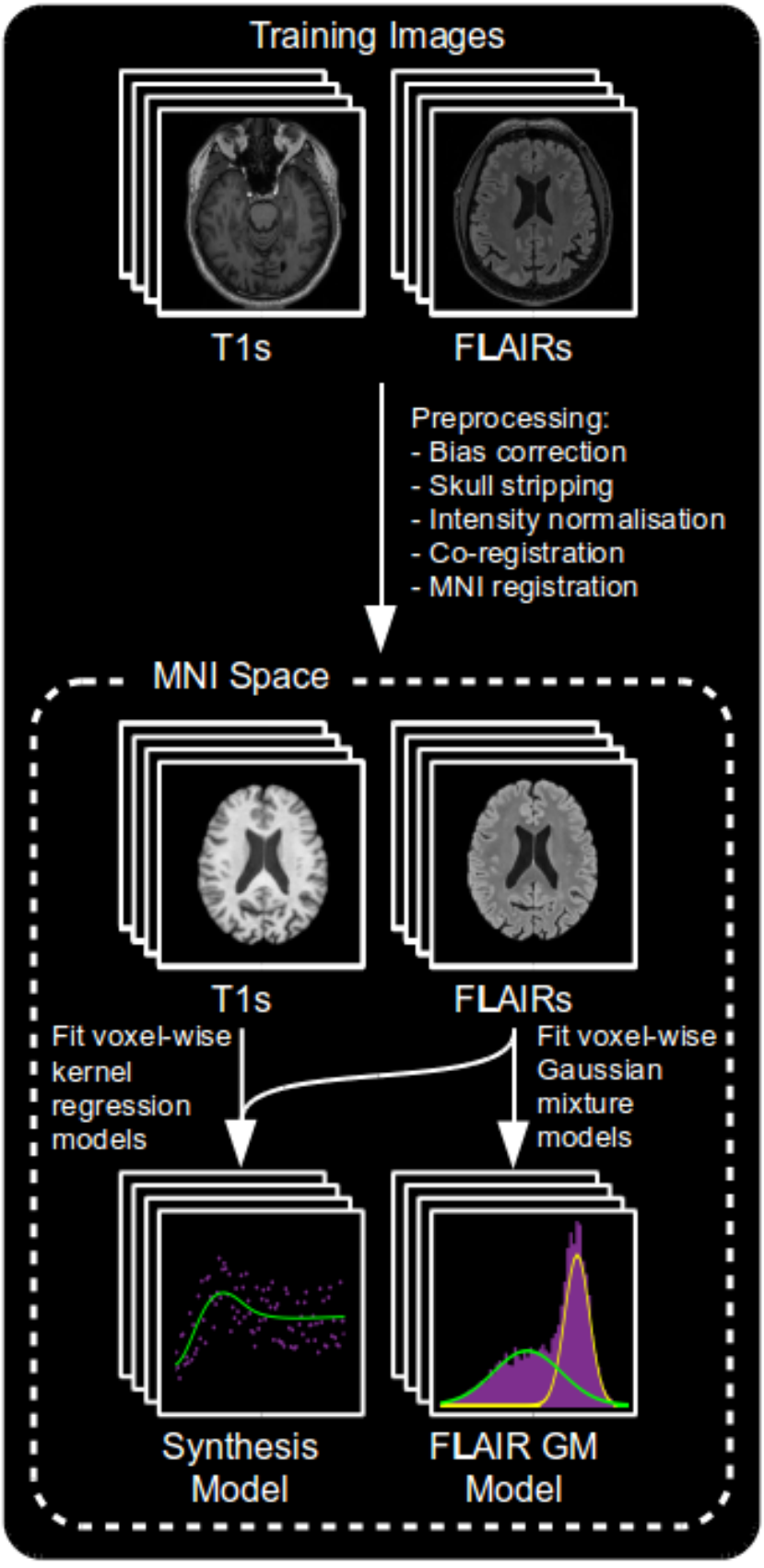


Figure 4.1: An overview of the training process.

20 and $a = 5$ were sufficient to produce useful images whilst remaining tractable ($<$ 24 hours to train, $<$ 1s to synthesise), with larger values having negligible impact on final results.

An example showing the models produced at two voxels is shown in Figure 4.2. The top right figure clearly displays the desired relationships in a location which can contain WM, GM or Cerebrospinal Fluid (CSF). The brightest T_1 intensities correspond to darker FLAIR intensities, corresponding to WM appearing brighter on T_1 images than on FLAIR. GM appears darker on T_1 images and brighter on FLAIR, explaining the peak of the model. Finally, the darkest T_1 intensities correspond to CSF, as is the case on FLAIR, which is represented by the leftmost section of the model. However, the top left figure shows the model formed in a location containing only WM, equivalent to the rightmost section of the previous model. Since there is no more information upon which to fit the model, the model extrapolates to predict the same FLAIR intensity across the whole range of T_1 intensities. This gives the model the desired ability to predict normal looking WM even in the presence of hypo-intense T_1 visible lesions, such as those in Figure 2.7. This provides a set of models which can be used to predict a FLAIR image corresponding to a T_1 image on a voxel-by-voxel basis.

A consequence of using kernel regression for synthesis is that the contrast between WM and GM in the synthetic image is reduced. This is due to the smoothing effect encouraging the model away from the extreme intensity values and towards the mean. For a given T_1 intensity, the predicted FLAIR intensity is a weighted average of a set of observed FLAIR intensities. Because of this, the predicted FLAIR intensity will necessarily lie between the maximum and minimum observed intensities. As a result, the very highest and lowest FLAIR intensities would never be synthesised. To correct this, an intensity transfer function is computed for each training subject by using histogram matching (as implemented in MATLAB function *imhistmatch* with 256 histogram bins) to match the intensity histogram of the synthesised image to that of the FLAIR image. The median of these transfer functions (Figure 4.3) is computed and used to correct all images, the effects of which can be seen in Figure 4.4. As can be seen in Figure 4.3, the primary effect of this transformation is to increase the contrast between the extremes, thereby countering the tendency towards the mean caused by the regression model.

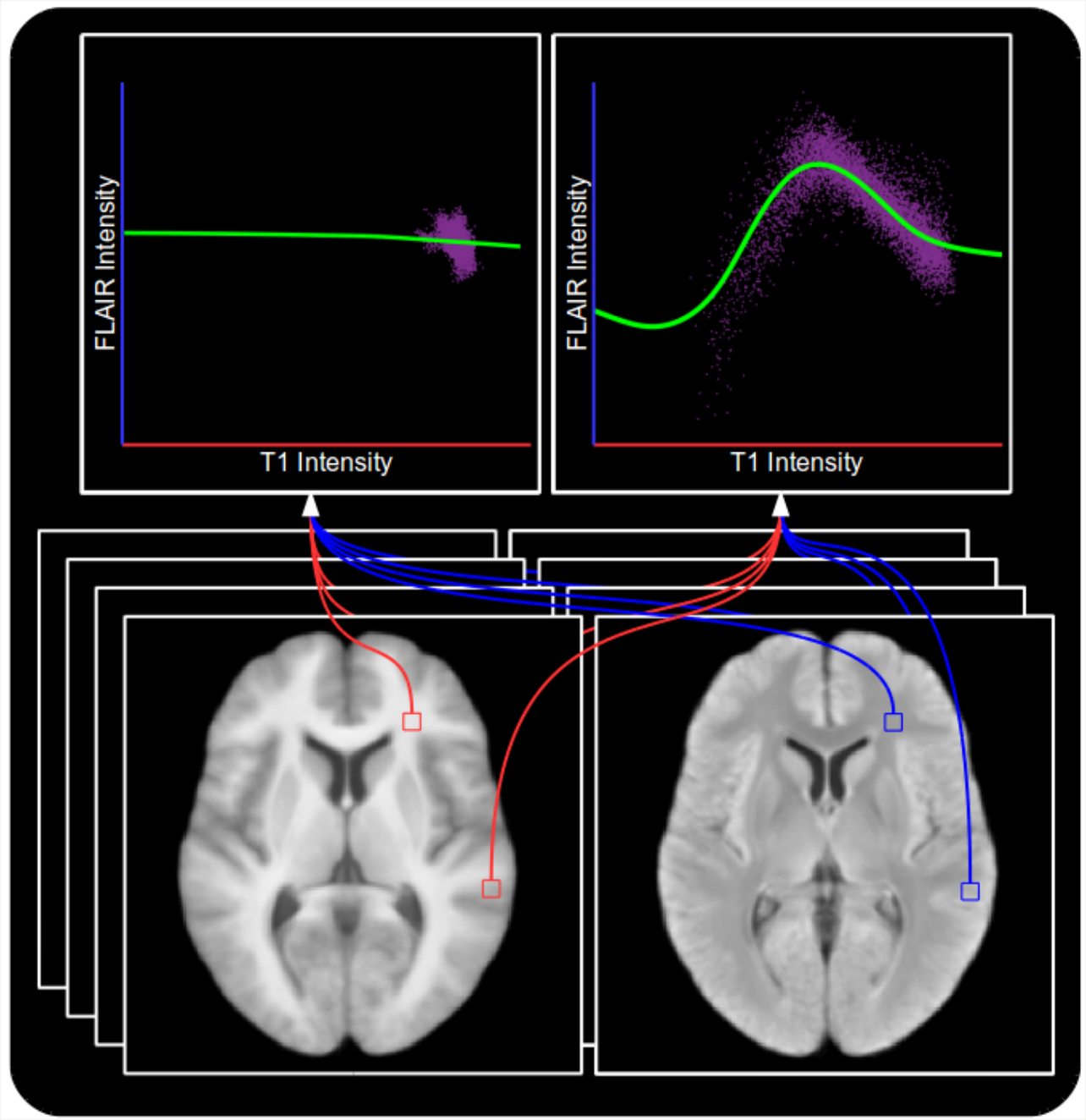


Figure 4.2: Two models produced using kernel regression to act as a mapping from T_1 to FLAIR intensities. Top left: A model produced at a location within the WM which contains only WM voxels. Top right: A model produced at a location which can contain WM, GM or CSF voxels. Bottom left: Mean T_1 training image. Bottom right: Mean FLAIR training image. Note that the model produced from WM, GM and CSF voxels is more complex than the one produced within the WM as a result of having to capture more intensity relationships, and that the extrapolation in the case of the latter provides the ability for the model to predict healthy WM FLAIR intensities even in the presence of T_1 visible pathology.

Gaussian Mixture model

$\mathbf{L}^{\mathbf{FLAIR}}$ is a representation of the likelihood of a voxel intensity being abnormal given previous knowledge of the expected distribution of intensities at each location. The distribution of

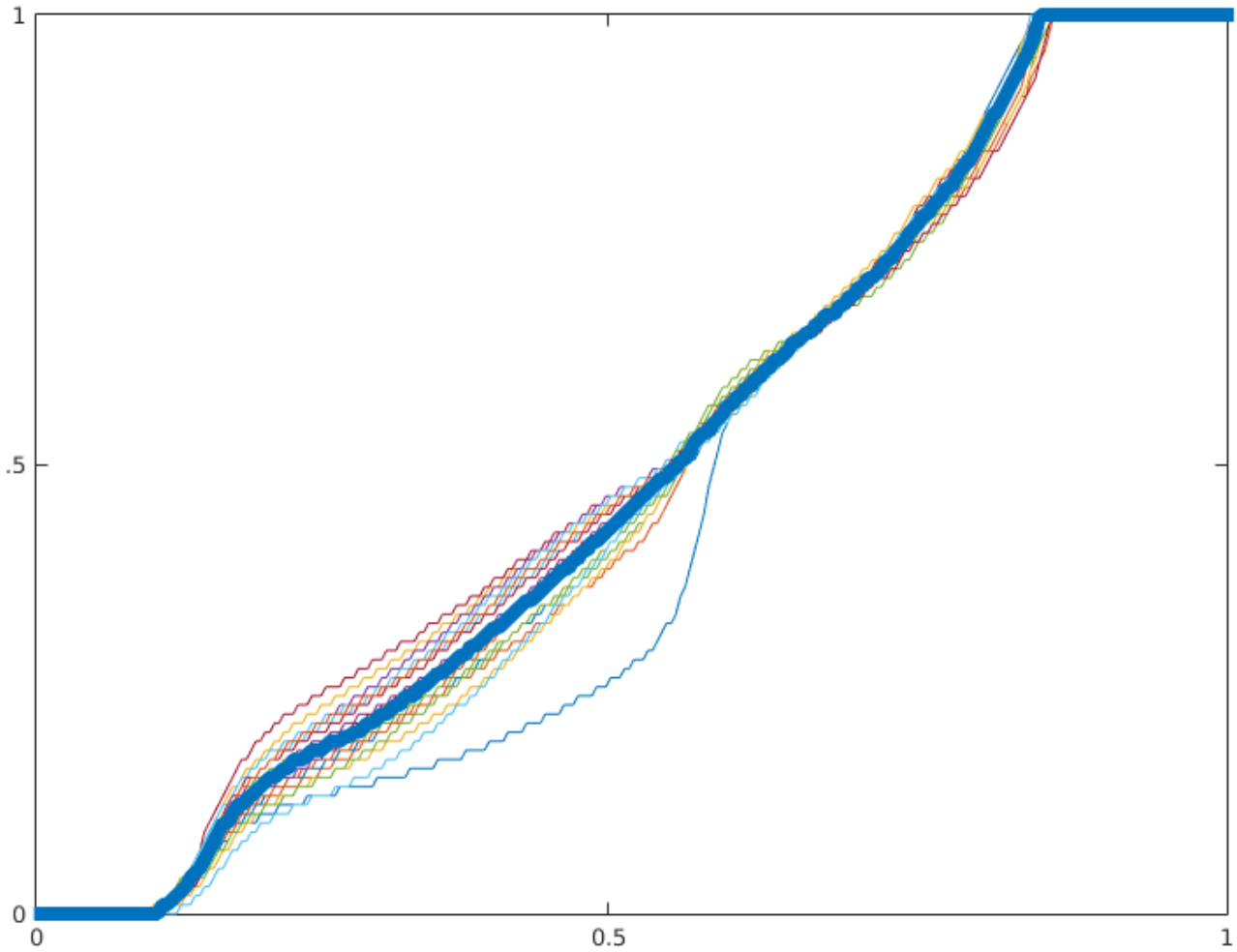


Figure 4.3: Transfer functions computed to map synthetic FLAIR images to their corresponding training FLAIR images. Thick blue line indicates the median which is used to correct all images.

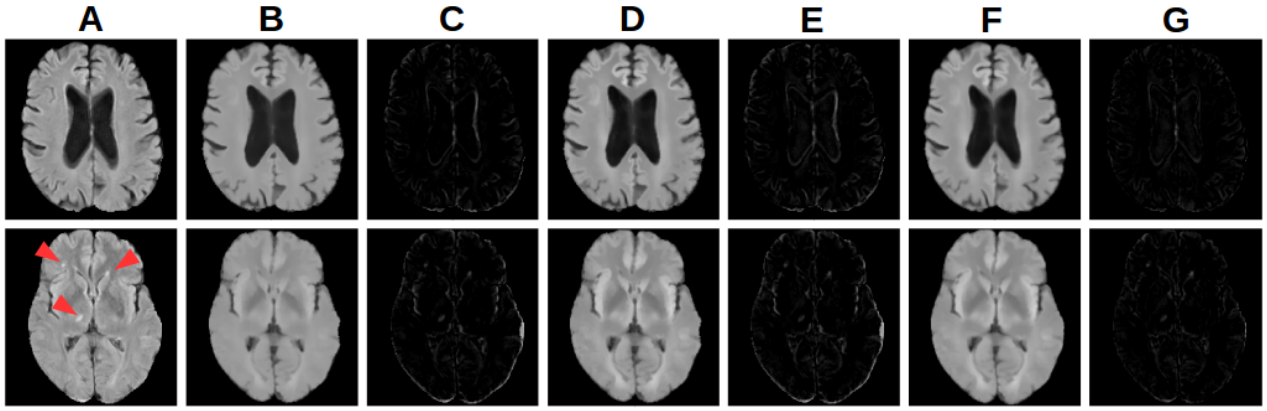


Figure 4.4: Effects of intensity correction and registration of synthetic images on a (top) pathology free and (bottom) pathological subject. (A) FLAIR image. (B) Rigidly registered synthetic image. (C) Difference image from (A) to (B). (D) Rigidly registered intensity corrected synthetic image. (E) Difference image from (A) to (D). (F) FFD registered intensity corrected synthetic image. (G) Difference image from (A) to (F). Note that the intensity correction and FFD registration do not prevent detection of the pathology (arrows).

intensities found across the whole brain is wide and complex, however at a voxel level, these distributions become narrower and easier to represent. It is common to treat intensities within a single tissue class as belonging to a Gaussian distribution, hence why many tissue segmentation algorithms are based upon an Expectation Maximisation (EM) framework [Zhang et al., 2001]. Intensities at a single voxel across a number of co-registered images will therefore likely belong to either one (when the voxel lies within a tissue class) or a mixture of two (when the voxel lies on the boundary between tissue classes) Gaussian distributions. An EM approach is therefore used [McLachlan et al., 2019] to learn a GMM with two components from $\mathbf{F}^{\mathbf{train}}$ at each voxel in MNI space. Due to a limited number of training images and the need for a lot of samples to confidently fit the GMM, voxels in a b-by-b-by-b patch around the target voxel are used, whilst boundary cases are handled by only considering non-zero intensities. Preliminary experiments showed that $b = 5$ provided sufficient information to confidently fit the models with 20 training images. An example showing the models produced at the same two locations as shown in Figure 4.2 is shown in Figure 4.5.

4.3.4 Testing

Having produced the two sets of models, they can now be applied to the test images to produce $\mathbf{L}^{\mathbf{SYN}}$ and $\mathbf{L}^{\mathbf{FLAIR}}$. A summary of the process can be seen in Figure 4.6.

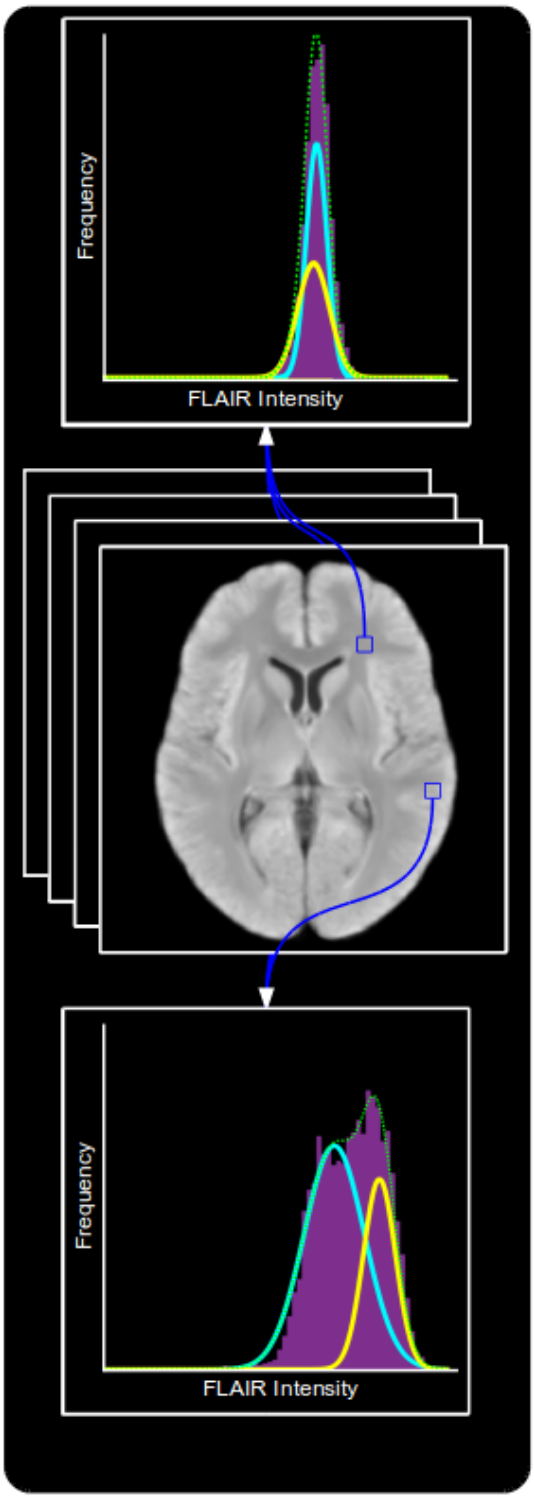


Figure 4.5: Two GMMs learned to represent the normal distribution of FLAIR intensities around their corresponding voxel. Top: A model produced at a location near the boarder between GM and WM. Middle: Mean FLAIR training image. Bottom: A model produced at a location within the WM. Note that the model produced from the border between WM and GM has two distinct components representing the two tissue types, whereas the model produced from within the WM contains two very similar components.

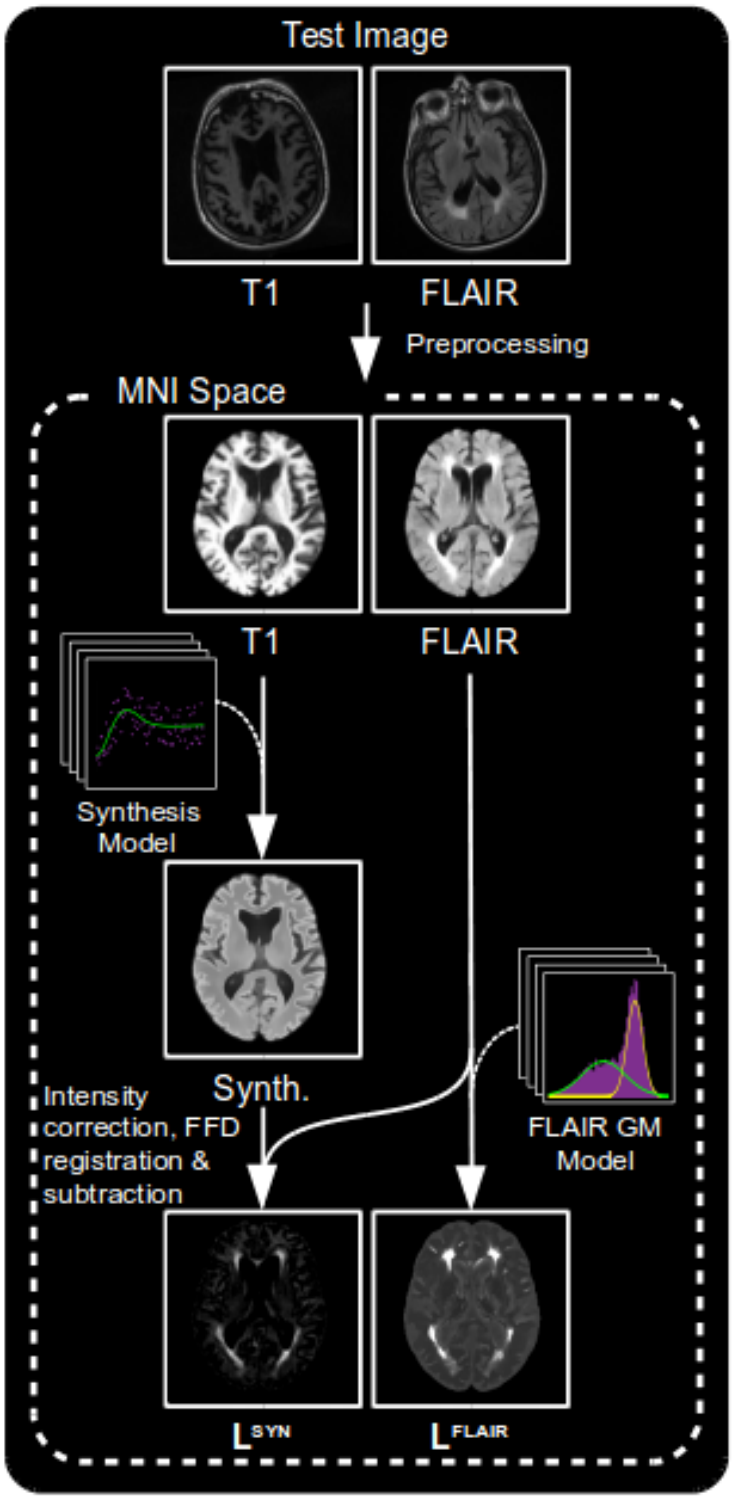


Figure 4.6: An overview of the process of creating the $\mathbf{L}^{\mathbf{SYN}}$ and $\mathbf{L}^{\mathbf{FLAIR}}$ likelihood maps.

$\mathbf{L^{SYN}}$

To synthesise a voxel $\mathbf{x}$ of synthetic image $\mathbf{S}$ using regression model M, the corresponding voxel in the subject's T_1 image, $\mathbf{T_x}$, is capped at t_{max} and turned into an index $i = \lceil m\mathbf{T_x}/t_{max} \rceil$. This index is then used to index into $\mathbf{M_x}$ to give $\mathbf{S_x}$. The intensities of $\mathbf{S}$ are finally adjusted using the previously computed transfer function. An example of successful pseudo-healthy synthesis in the presence of WMH can be seen in Figure 4.8.

As we will be performing voxel-wise comparisons of $\mathbf{F}$ and $\mathbf{S}$, it is important to have a good registration between them. As discussed earlier, studies have shown the benefits of using synthetic images to achieve more accurate multi-modal registrations by reducing the problem to a mono-modal one between the synthetic and target images. $\mathbf{S}$ is therefore registered directly to $\mathbf{F}$, producing $\mathbf{S^F}$. Despite this registration theoretically being rigid, a small non-linear term is introduced. This is to make the registration more robust to artefacts present in the either one of the images, in particular, distortions caused by eddy currents, and by partial volume effects often caused by FLAIR images having a large slice thickness.

A special case must be made for the region around the ventricles. Small hyper-intensities around the ventricular wall known as "bands" and "caps" are common in ageing and can be a result of several phenomena [Barkhof et al., 2011]. The presence of these bands and caps in the otherwise healthy training data leads to the undesired synthesis of clinically relevant WMH around the ventricles, see Figure 4.7. To avoid this leading to inaccurate segmentations, the intensities of WM in the synthetic images within 15 mm of the ventricles, as determined by a distance transform, are capped at the value corresponding to the expected intensity of healthy WM in this region.

$\mathbf{L^{SYN}}$ is then computed as $\mathbf{F} - \mathbf{S^F}$. At this point an approximate segmentation could be formed by applying a threshold to $\mathbf{L^{SYN}}$, however, there are situations which could cause errors to arise in the resulting segmentation. Artefacts in the T_1 image, particularly ringing artefacts, will cause errors in the synthesised image. These could introduce both false positives (seen in Figure 4.9), and false negatives should the ringing negate the signal from a lesion. Cortical

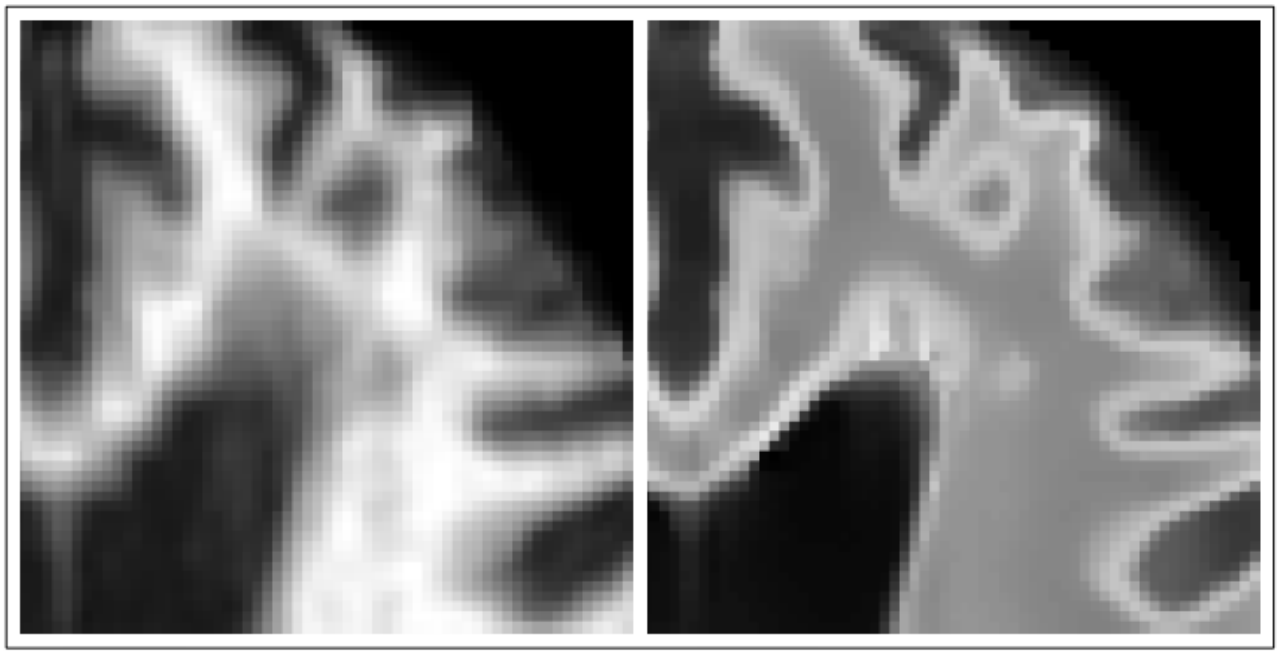


Figure 4.7: An example where periventricular WMH has been synthesised. Left: Normalised T_1 image. Right: Corresponding synthetic FLAIR image.

infarcts can sometimes be synthesised as hyper-intense as a result of being treated like GM due to their proximity to the cortex, seen in Figure 4.10. Whilst juxtacortical infarcts are brighter than normal GM on T_2-w images, the difference in intensity will be small, and could fall under a threshold. Finally, the high slice thickness common in FLAIR images can result in partial volume effects. These are particularly visible in the axial plane at the boundaries between CSF and WM or GM, such as at the top of the 3^{rd} and 4^{th} ventricles and the base of the frontal and temporal lobes. The synthetic image formed from the higher resolution T_1 image will not suffer these effects and will, therefore, appear brighter within the brain matter, leading to potential false positives.

In order to limit false positives due to T_1 artefacts and FLAIR partial volumes, and to reinforce areas of small differences in $\mathbf{L}^{\mathbf{SYN}}$ such as could be seen in the case of lesions in or near the cortex, additional information related to the brightness of the FLAIR image is required. This is obtained from $\mathbf{L}^{\mathbf{FLAIR}}$.

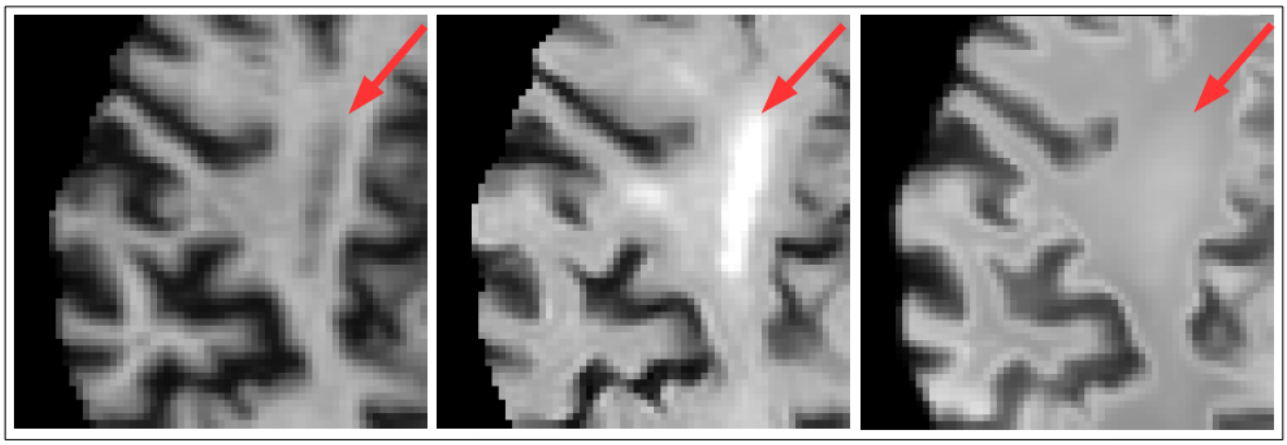


Figure 4.8: A case where a lesion is correctly synthesised as the same intensity as the surrounding WM. Left: T_1 image. Middle: FLAIR image. Right: Corresponding synthetic FLAIR image.

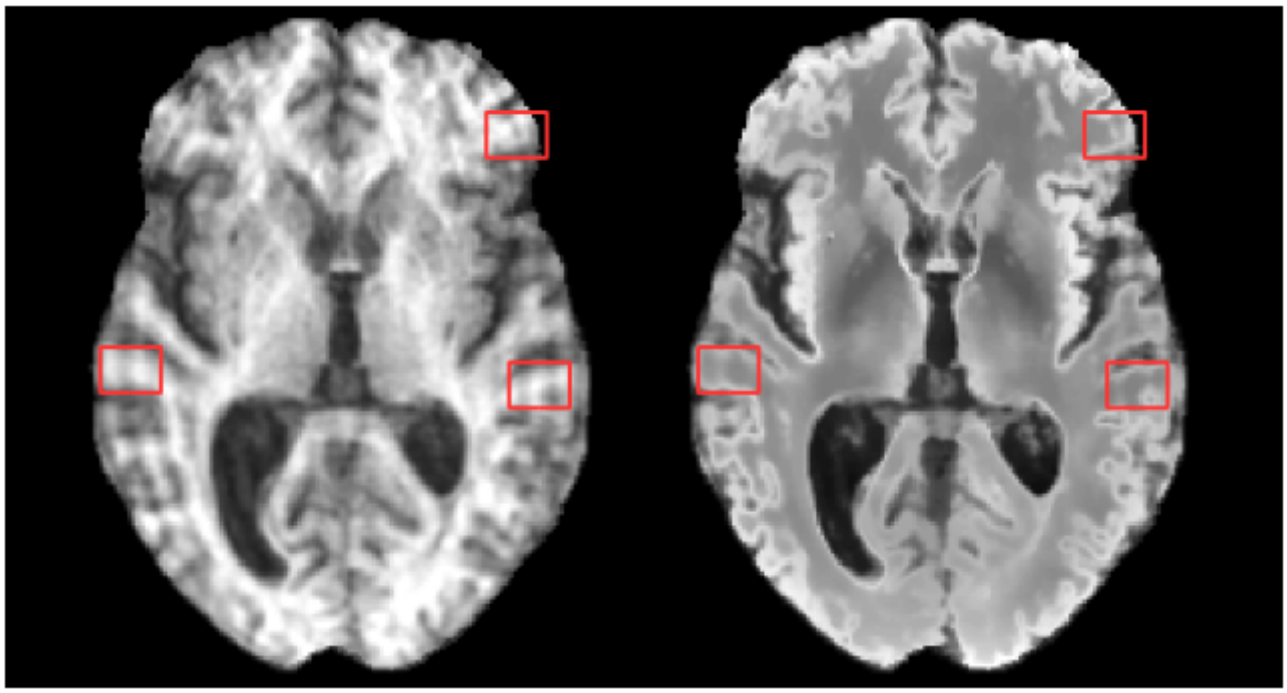


Figure 4.9: A case where ringing artefacts in a subject's T_1 image results in errors in the synthesised FLAIR image whereby juxtacortical WM is synthesised as GM in the indicated locations. Left: T_1 image. Right: Corresponding synthetic healthy FLAIR image.

$\mathbf{L}^{\mathbf{FLAIR}}$

To compute $\mathbf{L}^{\mathbf{FLAIR}}$, a relative likelihood is computed at each voxel reflecting the likelihood of that voxel being abnormal given the previously computed GMMs. To assign a likelihood to a given voxel, $\mathbf{x}$ in a test image, the log-likelihood of the intensity of the voxel is computed using the corresponding two- component GMM, parameterised by weights $(w_{1,\mathbf{x}}, w_{1,\mathbf{x}})$, means

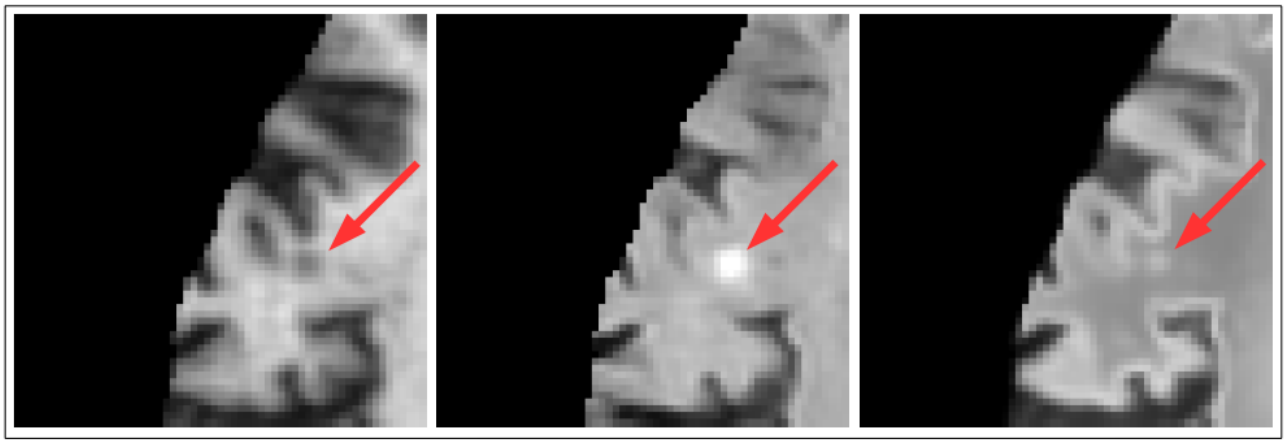


Figure 4.10: A case where a lesion close to the cortex is mistakenly synthesised as hyper-intense. Left: T_1 image. Middle: FLAIR image. Right: Corresponding synthetic FLAIR image.

($\mu_{1,\mathbf{x}}$,$\mu_{2,\mathbf{x}}$) and standard deviations ($\sigma_{1,\mathbf{x}}$,$\sigma_{2,\mathbf{x}}$). The resulting value will be large for both abnormally hyper- and hypo-intense voxels. To ensure only hyper-intense voxels are identified the likelihood is set to zero in regions with a FLAIR intensity $\mathbf{F_x}$ less than mean of the average intensities of GM and WM, previously set to 1000 during normalisation.

$$\mathbf{L}_{\mathbf{x}}^{\mathbf{FLAIR}} = \begin{cases} w_{1,\mathbf{x}} \frac{1}{\sqrt{2\sigma_{1,\mathbf{x}}^2 \pi}} e^{\frac{\mathbf{F_x} - \mu_{1,\mathbf{x}}^2}{2\sigma_{1,\mathbf{x}}^2}} + w_{2,\mathbf{x}} \frac{1}{\sqrt{2\sigma_{2,\mathbf{x}}^2 \pi}} e^{\frac{\mathbf{F_x} - \mu_{2,\mathbf{x}}^2}{2\sigma_{2,\mathbf{x}}^2}} & \text{if } \mathbf{F_x} \geq 1000 \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

4.3.5 Combining $\mathbf{L}^{\mathbf{SYN}}$ and $\mathbf{L}^{\mathbf{FLAIR}}$

To combine the information from $\mathbf{L}^{\mathbf{SYN}}$ and $\mathbf{L}^{\mathbf{FLAIR}}$, a similar framework to that proposed in [Karpate et al., 2015], where the authors combine a number of probability maps using a supervised Support Vector Machine (SVM), is used. Unsupervised one-class SVMs, such as in [Azami et al., 2016], were chosen to remove the need for labelled data and to maintain the proposed method's flexibility by allowing it to be used for general abnormality detection and not be restricted to a particular pathology present in a training set.

Training

The SVMs are trained using the same subjects which formed the training set used to train the models used to produce $\mathbf{L}^{\mathbf{SYN}}$ and $\mathbf{L}^{\mathbf{FLAIR}}$, with both likelihood maps in SVM space. A 3-by-1 feature vector is computed for each voxel containing the values of $\mathbf{L}^{\mathbf{SYN}}$, $\mathbf{L}^{\mathbf{FLAIR}}$ and an in house probabilistic WMH_{pvo} atlas generated by averaging co-registered manual WMH_{pvo} segmentations, a full description of which can be found in [Chen et al., 2015a].

A separate one-class SVM is trained for WM and GM. Fifty-thousand training points are randomly sampled from the feature vectors coming from each tissue class with an outlier percentage of 5% and 0.3% for the WM and GM classifiers respectively. These percentages were chosen empirically by visually assessing the resulting classifier's tendency to over/under-segment within each tissue class. Apparent over-segmentation lead to the outlier percentage being increased, whilst under-segmentation lead to a decrease.

Testing

To analyse a test image, the corresponding $\mathbf{L}^{\mathbf{SYN}}$ and $\mathbf{L}^{\mathbf{FLAIR}}$ likelihood maps are combined with the WMH_{pvo} atlas to form a feature vector at each voxel. Vectors are then classified using the previously trained one-class SVM corresponding to the tissue type which has the greater probability at that voxel. If the voxel falls outside of the decision boundary, and therefore considered an outlier, a score is formed for that vector defined as its distance from the decision boundary. A single likelihood map, $\mathbf{L}^{\mathbf{SVM}}$, is formed from these scores.

CRF refinement

To binarise and remove false positives from $\mathbf{L}^{\mathbf{SVM}}$ a final post processing step is applied, using a 3-dimensional (3D) fully connected CRF, described first in [Krähenbühl and Koltun, 2011] and extended to 3D and implemented in [Kamnitsas et al., 2017b].

Protocol	1	2	3
Number (test/train)	18/5	70/11	39/4
T_1 TR/TE/TI (ms)	9/440	9.7/3.984/500	
FLAIR TR/TE/ TI (ms)	9002/147/2200	9000/140/2200	
Ground Truth	Expert corrected histogram segmentation	Multispectral colour-fusion-based semi-automatic segmentation [1]	Expert corrected histogram segmentation
Lesion Types Present	WMH_{pvo}	WMH / Cortical infarcts	WMH_{pvo}

Table 4.1: Summary of the acquisition and segmentation protocols present in the dataset. [1][Valdes Hernandez et al., 2015, Valdés Hernández et al., 2013]

4.4 Experiments

To evaluate the performance of the proposed method, it is compared to three of the publicly available methods for lesion segmentation. Two methods from the LST, LST-LGA and LST-LPA, and LesionTOADS [3].

4.4.1 Data

The data for evaluation comes from the Edinburgh SVD dataset described in 2.6.2 and summarised in Table 4.1. The 20 subjects with the lowest lesion volume (so as to maximise healthy tissue) were selected to form $\mathbf{T}^{\mathbf{train}}$ and $\mathbf{F}^{\mathbf{train}}$ and excluded from further analysis. The manual masks for these subjects were dilated by one voxel and used to mask out regions of pathology from the training process. Note that this step would not be necessary if pathology free subjects were available to form the training set.

4.4.2 Evaluation metrics

A set of subject-wise similarity metrics were computed to quantify the performance of each method by comparing segmentation volumes V_a to target volumes V_t, and corresponding sur-

[3]Available at: `www.nitrc.org/projects/toads-cruise`

faces S_a and S_t. These include Dice Similarity Coefficient (DSC), Average Symmetric Surface Distance (ASSD), Hausdorff Distance (HD), Precision, Recall and Intra Class Correlation (ICC) as defined in Section 3.2.1. Scatter and Bland-Altman plots along with their associated metrics were also calculated, in addition to:

- *Correlation with Fazekas score*: Spearman's rank correlation coefficient calculated between $|V_a|/|V_{ic}|$ and a combined Fazekas score over all subjects, where $|V_{ic}|$ is a subject's intercranial volume mask. A Fazekas score is a clinical measure of WMH, comprising of two integers in the range $[0, 3]$ reflecting the degree of periventricular WMH and deep WMH respectively. For the purposes of this comparison, the two scores were added giving a single value in the range $[0, 6]$,

and two volume dependent metrics which provide additional insight into the conditions in which each method performs well, and where they are limited:

- *Lesion volume dependent DSC (DSC_l)*: The DSC calculated within the bounding box of each lesion, separated into groups corresponding to very small (< 0.01 ml), small (0.01-0.1 ml), medium (0.1-1 ml), large (1-10 ml), very large (> 10 ml) lesions. A lesion is defined as a single connected component within the reference segmentation. The bounding box of a lesion is defined as the smallest volume 3D box containing the lesion with dimensions parallel to the axes of the global coordinate system,
- *Subject volume dependent DSC (DSC_s)*: The DSC for subjects separated into groups corresponding to very low (< 5 ml), low (5-10 ml), medium (10-15 ml) and high (>15 ml) lesion volume according to reference segmentations.

4.4.3 Compared methods

The methods were selected for comparison, LST-LPA, LST-LGA and LesionTOADS, are summarised below. See Section 3.2.2 for further details.

- *LST-LGA*: One of the methods available in the LST. LST-LGA [Schmidt et al., 2012] is an unsupervised method which requires both a T_1 and a FLAIR image. The T_1 image is used to create a tissue type segmentation using an expectation maximisation approach. These tissue maps are propagated to the FLAIR image and used to create an initial lesion belief map which is binarised using a tunable threshold, κ. The authors suggest a κ value of 0.3, although they strongly encourage that this value is optimised for a particular dataset. The resulting segmentation is used as a seed for a region growing algorithm. The output of the algorithm is a probabilistic lesion map which must then be thresholded. Parameters (suggested): κ (0.3), threshold (0.5).

- *LST-LPA*: The second algorithm available in the LST. LST-LPA is a supervised algorithm which has been trained on 53 subjects with severe MS lesion patterns, and requires only a FLAIR image. A number of covariates for a logistic regression model are derived from the FLAIR image including a lesion belief map similar to the one produced by LST-LGA. The trained model is then used to assign a lesion probability estimate for each voxel, which is thresholded. Despite being supervised, the fact the model has been previously trained means it can be directly applied without requiring a training set. Parameters (suggested): threshold (0.5).

- *LesionTOADS* [Shiee et al., 2010]: This unsupervised algorithm introduces lesion segmentation to a previously developed structural segmentation method - TOpology preserving Anatomical Segmentation (TOADS) - by incorporating an additional lesion class. TOADS performs iterative segmentation driven by both statistical and topological atlases to ensure intensity and topological constraints are observed. LesionTOADS introduces a new class within the WM, with the union of the lesion and WM class following the same topological constraints as the original WM class. The algorithm requires both a T_1 and FLAIR image and outputs both a lesion and structural segmentation.

For each method, experiments were performed using both default parameters and optimised parameters based upon a grid search across one or two parameters which maximised DSC. For the proposed method these parameters relate to the CRF, with the default parameters being

those suggested in the CRF implementation [4] adjusted for an isotropic voxel grid. During optimisation, two parameters were varied. $w^{(2)}$ adjusts the relative weighting between the two CRF energy terms, and σ_γ determines how strongly homogeneity within the segmented region is enforced. Average subject-wise metrics and correlations for each method can be seen in Table 4.2, whilst volume dependent metrics for the optimal parameters can be seen in Tables 4.3 and 4.4. Significance testing at a 5% significance level was performed using Wilcoxon signed rank tests on subject wise metrics, and by comparing 95% confidence intervals for ICC.

Whilst LST-LPA and the proposed method successfully ran on all subjects, LST-LGA and LesionTOADS failed to run on two and three subjects respectively. Intercranial volume was also unavailable for two subjects. Results are given across all subjects for which the method was successful, whereas comparisons between methods were only taken across subjects which were successfully processed by both methods.

The results were also analysed by grouping subjects into the three acquisition protocols and computing the average DSC over each protocol, giving further insight into the strengths and weaknesses of each method, Table 4.5.

4.4.4 Clinical validation

In addition to the above quantitative evaluation, a clinical validation was also performed by examining the coefficients of a general linear model formed from the normalised segmentation volumes of each method and a number of clinical and radiological variables. These coefficients are then compared to those formed from a model relating the variables to the reference segmentations. The models are composed as such:

[4] Available at: `github.com/Kamnitsask/dense3dCrf`

$$\begin{aligned} Vol\%_i^{method} = &\beta_0 \ + \ \beta_1 \ Age_i \ + \ \beta_2 \ Gender_i \ + \ \beta_3 \ Diabetes_i \ + \\ &\beta_4 \ Hypertension_i + \ \beta_5 \ Hyperlipidaemia_i \ + \ \beta_6 \ Smoking_i \ + \\ &\beta_7 \ Cholesterol_i \ + \ \beta_8 \ PVSBG_i \ + \ \beta_9 \ DeepAtrophy_i + \epsilon_i, \end{aligned} \tag{4.3}$$

where $Vol\%_{method}$ is the lesion segmentation volume for each method as a percentage of intercranial volume, i indicates a particular subject, $Diabetes$, $Hypertension$ and $Hyperlipidaemia$ are binary variables, $Smoking$ is an integer (range [0,2] -never smoked, used to smoke, smokes), Cholesterol (mmol/L), PVS_{BG} is a radiological observation reflecting the number perivascular spaces in the basal ganglia [Potter et al., 2015], $DeepAtrophy$ is a radiological observation reflecting the degree of deep cortical atrophy [Farrell et al., 2009], ϵ is a residual error term and β is the set of coefficients which minimises $\sum_i \epsilon_i$. $Gender$ is included to remove bias but is not considered a risk factor and therefore not reported.

The strength of association between each clinical or radiological variable and the lesion volume produced by each method were measured by conducting a t-test for each coefficient β_i individually under the hypothesis that $\beta_i = 0$. The test statistic is found by dividing the coefficient estimate by its standard error computed during the fitting process. By setting a 5% significance level, the set of variables which have the strongest association with the measured lesion volume was found for each method.

An additional set of models were formed by replacing $PVSBG_i$ in Equation 4.3 with $Fazekas_i$, being the combined Fazekas score for subject i. Whilst expected to be strongly associated, comparing the β_8 coefficient calculated for each automated method to that calculated for the reference segmentations provides a further indicator as to which methods more accurately model the process of producing the reference segmentations.

Note that evaluation is carried out across only the subjects ($n = 96$) for which all clinical and radiological variables are available.

4.5 Results and Discussion

When comparing methods it is necessary to understand the aims and limitations of each algorithm. The methods contained in the LST were developed to segment MS lesions, while LesionTOADS aims to segment both WMH and MS lesions. These methods are therefore only interested in lesions within the WM, and restrict their search to reflect this using a WM tissue segmentation. On the other hand, the proposed method aims to segment all hyperintense lesions on FLAIR, including WMH, MS lesions and cortical infarcts, and as such, cannot restrict the search to the WM. Both approaches have advantages and disadvantages, which are reflected in the results and discussed in the following sections. The main advantage of restricting the search to the WM is that it avoids false positives occurring in the GM. This is important as GM can have a similar intensity distribution to WMH and MS lesions on FLAIR, and can therefore be a considerable source of false positives. The obvious drawback is that such methods will struggle to identify cortical infarcts. Figure 4.11 shows some example segmentations demonstrating the consequences of these approaches.

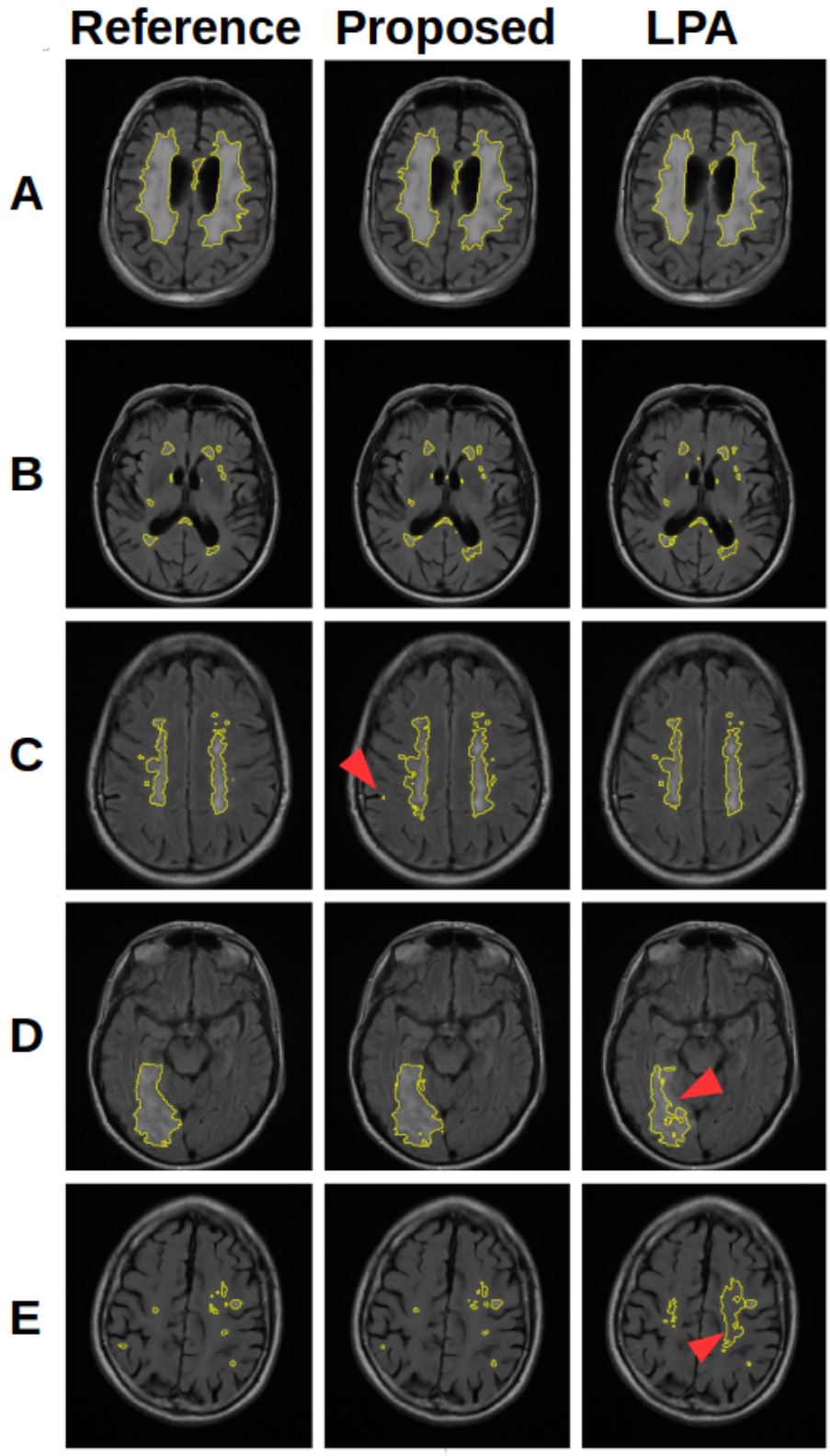


Figure 4.11: A selection of segmentations showing the features of the proposed method and LST-LPA. (A) and (B) show cases where both methods perform well. (C) shows a case where the proposed method produces false positive voxels (arrow) in the GM, not present in LST-LPA which does not consider GM. (D) shows a large infarct extending into the cortex where the extension into the cortex (arrow) is poorly segmented by LST-LPA. (E) shows a case where small lesions are missed by LST-LPA, despite considerable over segmentation (arrow).

Table 4.2: Table showing the results of each method over the whole dataset. Optimal parameter combinations (see 4.4.3) indicated by *. Statistical differences between the closest competitor (optimised LST-LPA) and the proposed method at a 5% significance level are bold. For comparison, correlation between ground truth volumes and Fazekas scores is 0.829.

Method	Parameters	DSC	ASSD	HD	Prec.	Recall	ICC	Faz. Corr.
LST-LGA	$\kappa = 0.3$ $t = 0.5$	0.382	5.77	48.6	0.925	0.265	0.693	0.782
LST-LPA	$t = 0.5$	0.536	2.60	37.3	0.926	0.416	0.874	0.846
LesionTOADS		0.497	2.74	34.3	0.667	0.498	0.488	0.358
LST-LGA	$\kappa = 0.11$* $t = 0.01$*	0.473	4.54	39.9	0.698	0.403	0.836	0.767
LST-LPA	$t = 0.15$*	0.683	1.62	**33.3**	0.759	0.681	0.952	0.805
Proposed	$w^{(2)} = 8$* $\sigma_\gamma = 2.5$*	**0.703**	**1.23**	38.6	0.763	0.695	**0.985**	0.862

4.5.1 Whole dataset analysis

When considering the dataset as a whole, Table 4.2 shows that the proposed method generally outperforms the existing methods, with significant improvements in DSC, ASSD and ICC. Despite being developed for and trained on MS lesions, LST-LPA performs very well, and is the closest competitor across these metrics, with a significantly superior HD. This superior HD can be explained by the reduced likelihood of false positives in the GM when compared to the proposed method, as discussed earlier. Any tendency towards false positives far away from real lesions, such as in the GM, will be strongly punished by HD. LesionTOADS and LST-LGA both fall well short of LST-LPA and the proposed method. It is clear that the suggested thresholds of 0.5 and κ of 0.3 result in considerable under segmentation and overall poor results. It is however interesting to observe that these methods do achieve high correlations with Fazekas scores despite lower performance compared to ground truth segmentation. This suggests that a fully accurate segmentation may not be necessary to predict a Fazekas score. The proposed method has the strongest correlation with Fazekas scores (0.862), which is stronger that of the reference segmentations (0.829), though with a p-value of 0.18, it is not possible to say conclusively that the automated method outperformed the reference segmentations in this regard. Similarly, the power (56%, non-parametrically estimated through bootstrapping) of the DSC comparison between LST-LPA and the proposed method suggests that additional data would help to strengthen these conclusions.

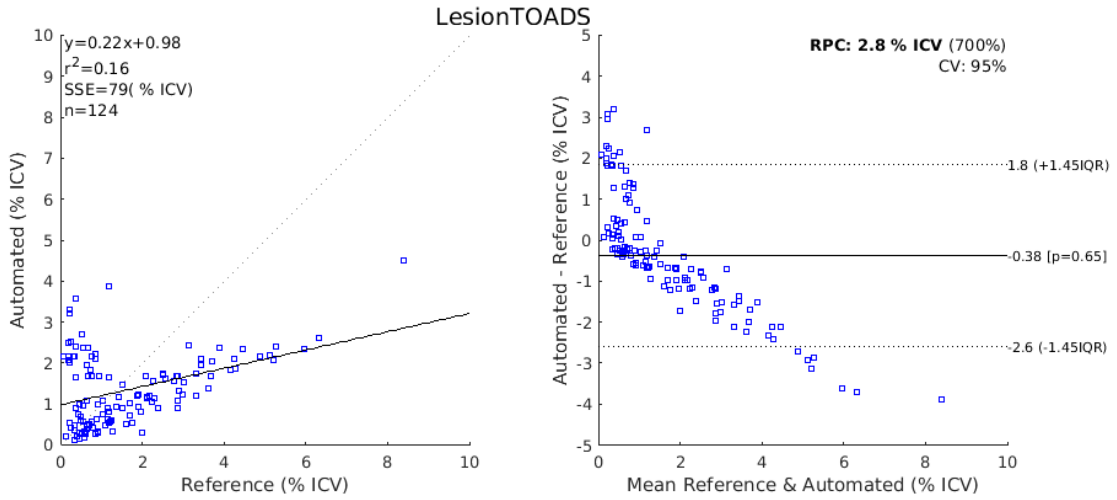


Figure 4.12: Scatter and Bland-Altman plots comparing of the lesion volumes (as a percentage of intercranial volume) from LesionTOADS to those from the reference segmentations.

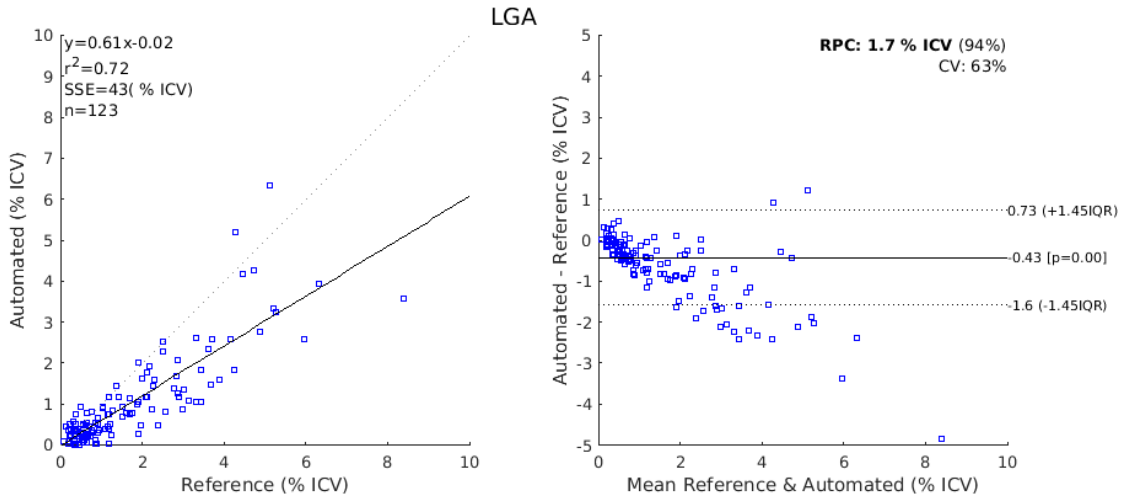


Figure 4.13: Scatter and Bland-Altman plots comparing of the lesion volumes (as a percentage of intercranial volume) from LST-LGA to those from the reference segmentations.

The relative performance of each method compared to one another indicated by these results are further supported by the scatter and Bland-Altman plots shown in Figures 4.12 to 4.15. We see a clear visual improvement going from LesionTOADS to LST-LGA, to LST-LPA, and to the proposed method, along with an improvement in the associated metrics. A common feature of LesionTOADS, LST-LPA and LST-LGA is a tendency to underestimate lesion volumes at larger lesion loads, whilst the proposed method appears unaffected. One contributory factor towards this could be the intensity normalisation procedure which was chosen so as to be unaffected by lesion load.

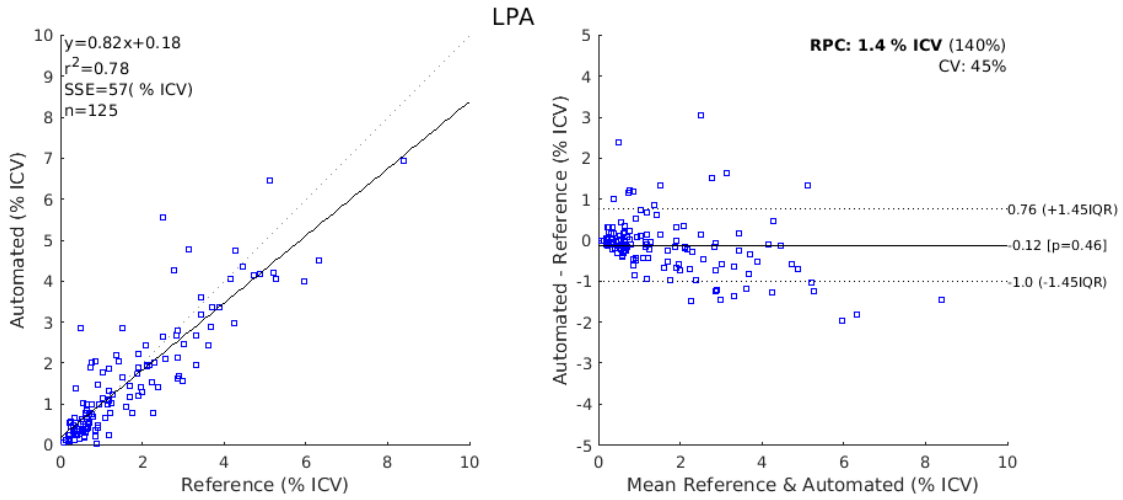


Figure 4.14: Scatter and Bland-Altman plots comparing of the lesion volumes (as a percentage of intercranial volume) from LST-LPA to those from the reference segmentations.

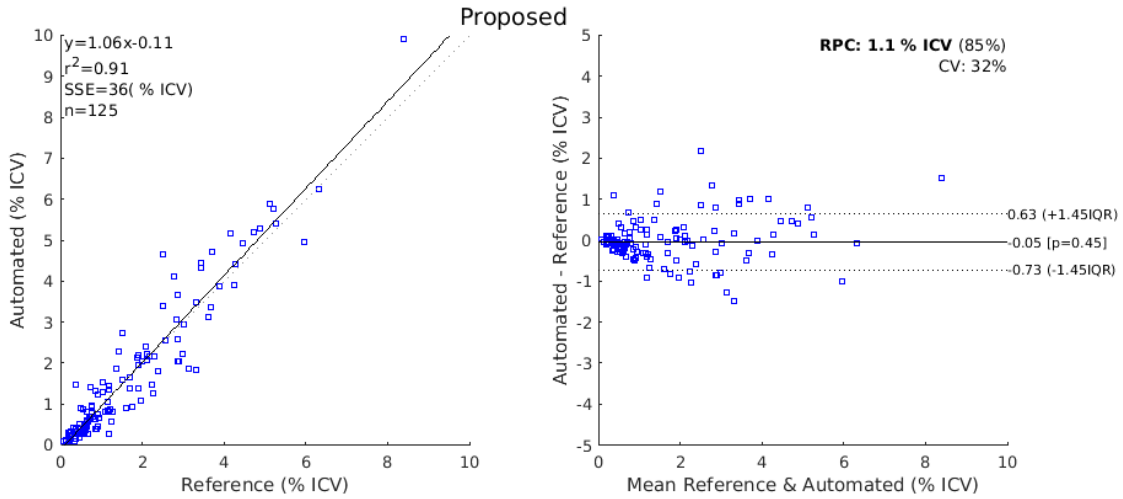


Figure 4.15: Scatter and Bland-Altman plots comparing of the lesion volumes (as a percentage of intercranial volume) from the proposed method to those from the reference segmentations.

4.5.2 Volume specific analysis

Table 4.3: Lesion volume dependent DSC (DSC_l) for each optimised method. Statistical differences between the closest competitor (optimised LST-LPA) and the proposed method at a 5% significance level are bold.

Method	<0.01 ml	0.01-0.1 ml	0.1-1 ml	1-10 ml	>10 ml
LesionTOADS	0.077	0.155	0.333	0.514	0.629
LST-LGA	0.024	0.048	0.214	0.467	0.599
LST-LPA	0.094	0.198	0.496	0.691	0.797
Proposed	**0.150**	**0.335**	**0.577**	**0.713**	0.807

Table 4.4: Subject volume dependent DSC (DSC_s) for each optimised method. While the proposed method obtains the largest DSC_s values, the differences with the closest competitor (optimised LST-LPA) are not significant

Method	<5 ml	5-10 ml	10-15 ml	>15 ml
LesionTOADS	0.157	0.440	0.426	0.614
LST-LGA	0.343	0.334	0.374	0.577
LST-LPA	0.558	0.615	0.569	0.762
Proposed	0.576	0.628	0.666	0.770

When the dataset is divided into subsets with different lesion volumes we see that the proposed method performs better across all subsets. Whilst individually not significant at a 5% level due to the lower power of the subsets, the consistency of these results leads to the significantly higher DSC observed in Table 4.2. We also observe the trend that DSC increases as lesion volume increases, shown in Table 4.4. This is an expected result and one which has been frequently observed [Griffanti et al., 2016]. A similar trend is observed when examining results on individual lesions in Table 4.3. The smaller the lesion, the lower the expected DSC. This is a feature of DSC and can be explained by a number of factors. First, the larger lesions present in subjects with a high total volume of lesions have a higher ratio of internal to boundary voxels. Internal voxels tend to be more hyperintense and have more support from adjacent voxels, leading to easier segmentation. Secondly, smaller lesions tend to be less hyperintense, reducing the contrast with surrounding tissue, making them harder to segment. Finally if we assume a rate of false positives due to noise or artefacts independent of total lesion volume, these will have a much larger impact on the DSC for subjects with a low total lesion volume than those with a high total lesion volume where the potential for true positives to counter the effects of the false positives is much greater.

A consequence of the above is that the overall DSC reported in Table 4.2 is dominated by the ability of the algorithm to detect large lesions. Over 80% of the total volume of lesions belong to lesions with a size >1 ml, and over 95% belong to lesions > 0.1 ml. However, small but strategically placed lesions can be clinically vital and the ability to detect these should form part of the evaluation of an algorithm. The results in Table 4.3 allow us to compare the performance of each method on differing sizes of lesion. We observe that whilst the proposed

Table 4.5: Table comparing average DSC for each method on images belonging to each protocol. Statistical differences between the closest competitor (optimised LST-LPA) and the proposed method at a 5% significance level are bold.

Protocol	1	2	3
LesionTOADS	0.431	0.535	0.445
LST-LGA	0.322	0.453	0.568
LST-LPA	0.688	0.678	0.690
Proposed	0.645	**0.710**	**0.719**

method and LST-LPA get similar results on the larger lesions, the proposed method performs much better than the other methods at detecting smaller lesions.

4.5.3 Protocol specific analysis

It is possible to gain further insight into the merits of each method by looking at the results over each of the three protocols present in the dataset, allowing for more direct comparisons between the methods. It is important to remember that subdividing the dataset in this way leads to a loss of sample power. Whilst the lower sample size is offset by stronger differences between LST-LPA and the proposed method in the cases of protocols 2 and 3 (power = 74% and 57% respectively), these are still lower than desired and the small sample size for protocol 1 leads to a power of just 2%. As such the results should only be considered along with other factors, such as algorithm design, to lend support to hypotheses regarding the strengths and weaknesses of each method.

Images acquired under protocols 1 and 3 contain only $\mathrm{WMH}_{\mathrm{pvo}}$ and are therefore ideal cases for both LST methods and LesionTOADS due to the lack of cortical infarcts. On the other hand, images acquired under protocol 2 can contain both WMH and cortical infarcts, the latter being more likely to be segmented by the proposed method. The results in Table 4.5 suggest that LST-LPA performs better on protocol 3 than on protocol 2, whilst the metrics for the proposed method are similar between the two protocols. This supports the hypothesis that LST-LPA suffers in the presence of cortical infarcts.

Protocol 3 allows for a direct and fair comparison between the methods, as it does not contain

cortical infarcts and is therefore not biased against the methods which only search in WM. Despite this, the proposed method significantly outperforms the other methods on protocol 3, indicating the superior results seen across the full dataset are not simply due to the ability to detect cortical infarcts.

However, both the proposed method and LST-LGA perform worse on protocol 1 than protocols 2 and 3, whereas LST-LPA performs equally well on protocols 1 and 3. Whilst the power of the comparison is extremely small, there are compelling reasons why LST-LGA and the proposed method might not perform as well on protocol 1 as protocol 3. While LST-LPA uses only the subject's FLAIR image, LST-LGA and the proposed method use both T_1 and FLAIR. The FLAIR acquisition protocol differs very little across the three protocols, however, the T_1 acquisition does. The T_1 images acquired under protocol 1 come from a spoiled gradient echo sequence, as opposed to the magnetisation prepared fast gradient echo sequence used in protocols 2 and 3. This leads to a lower contrast T_1 images in protocol 1, and a negative effect on the results of the two methods which use T_1 images.

Finally, recent work [Haller et al., 2016] has shown that protocol-specific MR parameters can systematically bias the results of automated volume estimation of a number of brain structures by 4-5%. We must, therefore, consider the possibility of a similar effect being present when estimating lesion volume. Whilst this is hard to observe from the results, given that the three protocols differ by more than just MR parameters, it should be considered as a potential contributory factor to explain the differences between the results from protocol 1 and those from 2 and 3.

4.5.4 Clinical Validation

Looking for associations between clinical and radiological measurements and calculated lesion volumes provides an alternative way to compare methods. Whilst the dataset used contains a variety of pathologies and degrees of abnormality, and as such strong associations with all risk factors are not expected, comparing what associations are found to those found using

Table 4.6: P-Values of the coefficients found using the model shown in Equation 4.3 . Bold indicates statistical significance of the coefficients from 0 at a 5% level.

WMH	Reference	LST-LGA	LST-LPA	Proposed	LesionTOADS
Age	0.82	0.88	0.11	0.55	$\mathbf{5 \times 10^{-3}}$
Diabetes	**0.03**	0.45	**0.01**	**0.02**	0.71
Hypertension	0.28	0.09	0.22	0.39	0.11
Hyperlipidaemia	0.37	0.87	0.24	0.29	0.78
Smoking	0.63	0.27	0.40	0.27	0.78
Cholesterol	0.95	**0.04**	0.12	0.11	0.53
PVSBG	$\mathbf{4 \times 10^{-13}}$	$\mathbf{7 \times 10^{-7}}$	$\mathbf{2 \times 10^{-9}}$	$\mathbf{2 \times 10^{-8}}$	0.13
DeepAtrophy	**0.02**	$\mathbf{2 \times 10^{-5}}$	$\mathbf{3 \times 10^{-5}}$	$\mathbf{6 \times 10^{-4}}$	0.40

Table 4.7: Coefficients of found using the model show in Equation 4.3 with $Fazekas$ in place of $PVSBG$. Bold indicates coefficients which are significantly different from 0 at a 5% level.

WMH	Reference	LST-LGA	LST-LPA	Proposed	LesionTOADS
Age	-3×10^{-4}	-5×10^{-4}	**0.019**	0.008	**0.032**
Diabetes	0.189	-0.083	**0.477**	0.248	-0.119
Hypertension	0.251	0.196	0.028	0.121	-0.118
Hyperlipidaemia	-0.098	-0.065	-0.029	0.043	-0.103
Smoking	-0.030	0.029	0.014	0.033	0.003
Cholesterol	0.107	-0.026	0.017	0.028	0.077
Fazekas	**0.649**	**0.320**	**0.555**	**0.717**	**0.150**
DeepAtrophy	**0.002**	**0.012**	**0.013**	**0.010**	**-0.002**

the reference segmentations provides confirmation that the methods being compared produce segmentations with the same distribution across subjects.

Figure 4.6 shows that there is a strong association between the reference segmentation volumes and perivascular spaces in the basal ganglia, deep atrophy and diabetes. This pattern is reflected in the results from LST-LPA and the proposed method, suggesting a good correspondence between these segmentations and the reference. The results from LST-LGA agree with two out of the three associations but also suggests an association with cholesterol which is not present in the reference. The results from LesionTOADS find only an association with age, sharing no associations with that of the reference. These results are in keeping with the previous observations, reinforcing the belief that LST-LPA and the proposed method both produce more accurate segmentations than the other two.

The coefficients in Figure 4.7 suggest that an increase in 1 in the combined Fazekas score

is associated with an increase in reference lesion volume of 0.649. This association is most similar to that found using segmentations from the proposed method (0.717), with those from LST-LPA (0.555) also similar. Again, LST-LGA is next closest, followed by LesionTOADS.

4.6 Conclusion

This chapter has presented a method for brain lesion segmentation through the use of a modality transformation algorithm, regardless of underlying pathology. It has shown that an apparently healthy FLAIR image can be synthesised from a subject's T_1 image and that the differences between this synthetic FLAIR and the real FLAIR can be combined with information from the real FLAIR to indicate the location of lesions. The resulting segmentations are objectively superior to a number of established methods across a range of clinically relevant metrics, including a particularly strong ability to detect smaller lesions. The results allow us to make the following conclusions:

- The proposed method significantly outperforms the existing methods on a heterogeneous dataset across most metrics.
- The proposed method does particularly well in cases with cortical infarcts, which are undetected by other methods.
- One of the biggest advantages of the proposed method is its ability to detect smaller lesions, something which, depending on the application, could be clinically highly relevant.
- Whilst not catastrophic, a limitation of the proposed method is that it requires both FLAIR and T_1 images, and any significant changes in T_1 acquisition protocols may negatively impact performance, see Table 4.5.

Future work will involve extending the framework to allow for the detection of unexpected hypointensities, such as lacunar cavities, and other hallmarks of SVD such as microbleeds and enlarged perivascular spaces. Modifying the approach to more readily handle a variety of

acquisition protocols, either through sequence normalisation [Roy et al., 2013] or through an extension of the regression model itself, will also make the method more robust.

Additionally, we mention in Section 4.3.4 that we must make a special case at the boundary between the top of the ventricles and the WM due to the impact of partial volume effects being particularly obvious in this area. This is one way that we reduce the impact of the high slice thickness FLAIR data, however, there are other areas where this could be further addressed, potentially leading to improvements. When working with low resolution or high slice thickness data, it is important to avoid resampling the image where possible, as doing so can degrade the image and lead to unrealistic appearances as the information from one large voxel is distributed over a set of smaller voxels. In this work we do this by synthesising the synthetic FLAIR in MNI space and then transform this into the FLAIR space, meaning the original FLAIR image is never resampled. However, in the current formulation, each training FLAIR image must be sampled onto MNI space to build the synthesis model. Future work will, therefore, investigate methods to avoid the need for this transformation. In addition, the process of resampling the higher resolution synthetic FLAIR image to the lower resolution real FLAIR image could also potentially be improved by considering the voxel size so as to better replicate a single large voxel in the real FLAIR image as a weighted average of multiple smaller voxels in the synthetic FLAIR image, as opposed to sampling a single slice through the image volume.

Future work will also involve investigating the effects of the various pre-processing steps used in the proposed approach. In particular, bias field correction algorithms can be affected by large regions of pathology, especially WMH, and removing this step could potentially lead to improvements. Additionally, while the choice of intensity normalisation procedure was carefully chosen, measuring the impact of using alternative methods could highlight flaws or provide additional confidence in the chosen method.

We have presented an approach to lesion segmentation motivated by various observations regarding the appearance of pathology in different MR imaging protocols. By doing so, we were able to largely avoid the need for labelled training data. However, supervised methods are frequently shown to lead to more accurate segmentations, particularly those which

utilise deep learning approaches. Work carried out in parallel to that described here (available in [Guerrero et al., 2018]) demonstrated how a supervised neural network can achieve similar performance to the method proposed in this chapter on the same data, while also differentiating between different lesion types. However, this work also highlighted one of the drawbacks of such approaches: the need for large amounts of accurately labelled training data. The following two chapters investigate how image synthesis can be used to alleviate some of these drawbacks and improve the performance of supervised deep learning approaches.

Chapter 5

GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks

5.1 Introduction

One of the biggest issues facing the use of machine learning in medical imaging is the lack of availability of large, labelled datasets. The annotation of medical images is not only expensive and time-consuming but also highly dependent on the availability of expert observers. The limited amount of training data can limit the performance of supervised machine learning algorithms which often need very large quantities of data on which to train to avoid overfitting. It is therefore of paramount importance to extract as much information as possible from what data is available.

Data augmentation is commonly used by many deep learning approaches in the presence of limited training data. Increasing the number of training examples through the rotation, reflection, cropping, translation and scaling of existing images is common practice during the training of learning algorithms, as it allows for the number of samples in a dataset to be increased by

factors of thousands [Krizhevsky et al., 2012]. Populating the training data with realistic, if synthetic, data in this way can significantly reduce overfitting and thus not only improve the accuracy but also the generalisation ability of deep learning approaches. This is of particular importance in convolutional neural networks which cannot easily learn rotationally invariant features unless there are a sufficient amount of examples at different angles in the training data.

Generative Adversarial Networks (GANs) offer a novel way to unlock additional information from a dataset by generating synthetic samples with the appearance of real images. This chapter proposes to use a GAN to model the underlying distribution of training data to allow for additional synthetic data to be sampled and used to augment the real training data in two brain segmentation tasks.

First proposed in [Goodfellow et al., 2014], GANs are a class of neural networks which aim to learn to generate synthetic samples with the same characteristics as a given training distribution. In the case of images, this involves learning to produce images (via a generator) which are visually so similar to a set of real images so that an adversary (the discriminator) cannot distinguish between them. The original formulation has since been built on to address problems such as training stability [Radford et al., 2015], low resolution [Berthelot et al., 2017, Zhang et al., 2017, Karras et al., 2017] and the absence of a true image quality based loss function [Arjovsky et al., 2017], and applied to tasks such as super resolution [Ledig et al., 2016], domain adaptation [Yoo et al., 2016], and reconstructing images from a minimal amount of data [Yeh et al., 2016]. See Section 3.1.1 for further details.

Various methods for using GANs to expand training datasets have been recently proposed. In [Shrivastava et al., 2016], the authors use an adversarial network to improve the quality of simulated images and use these for further training. In [Antoniou et al., 2017], the authors train a conditional GAN on unlabelled data to generate alternative versions of a given real image, and in [Zhu et al., 2017b], the authors use a CycleGAN to impose emotions on neutral faces to expand underrepresented classes. However, the use of non-conditional GANs to augment training data directly as a preprocessing step with no additional data has only very recently been explored [Amitai and Goldberger, 2018, Moradi et al., 2018], with promising results in

medical image classification tasks.

5.2 Motivation

The general procedure when faced with a dataset to be used in a machine learning task is to consider what sources of variance there are in the dataset and whether these are relevant to the features which the user hopes to identify. These will fall into two categories, *pertinent* and *non-pertinent* variance. Pertinent variance describes those features which are important to whatever information the user wishes to extract. In medical imaging, these are often features such as the size, shape, intensity and location of key components such as organs or lesions. Non-pertinent variance describes the features which vary between images which are not related to the important information the user wishes to extract. Examples of these are global intensity differences, position within the image field of view and appearance of unrelated anatomy. Exactly which sources of variance are pertinent or non-pertinent will depend on the application, and may not be known apriori. For example, in neuroimaging, whether a lesion is in the left or right hemisphere may or may not be diagnostically relevant, and so could fall into either category. There are therefore no rules which will fit every situation, and as such, considering and categorising each sort of variance is an important step requiring domain-specific knowledge.

Once the non-pertinent sources of variance have been identified, a decision is usually made on how to address them. Keeping too much non-pertinent variance in the final dataset can not only occlude the diagnostically important information but also lead to overfitting, especially in the relatively small datasets often used in medical imaging, where the trained model may learn to base its decision on coincidental correlations with irrelevant features. This is especially the case in deep learning methods, where features are learned from the data, while more traditional "handcrafted" features avoid much of this problem.

On the other hand, should non-pertinent variance be removed from the training set, it must also be able to be removed from any test instances. Common methods to remove such information include image registration to a standard space, intensity normalisation and cropping to

a region of interest. These are powerful tools to remove a lot of non-pertinent variance, substantially simplifying the training data distribution, and importantly, can easily be applied to test instances. For example, it is common in neuroimaging to perform intensity normalisation (both global and local bias field correction), co-registration and brain extraction.

The alternative to removing non-pertinent variance is data augmentation. One of the goals of data augmentation is to populate the data with a large amount of synthetic data in the directions of these non-pertinent sources of variance. The aim of this is to reduce this variance to noise, removing any coincidental correlation with labels and preventing its use as a discriminative feature. An example of this would be to augment using rotations of real images in cases where orientation is irrelevant to the desired model output.

There are however some sources of non-pertinent variance which can neither be removed or augmented. For example, patient-specific variation in non-relevant anatomy. It may not be possible to remove this anatomy through cropping or to define an accurate enough model to augment this variance with realistic cases. Previous efforts [Krivov et al., 2017] have been made to alleviate this particular source of non-pertinent variance by propagating lesions to healthy brains, thereby incorporating additional examples of non-pertinent anatomical variance. Despite this, there will always be sources of non-pertinent variance which cannot be removed or augmented, for which the chosen model must account for.

The other use of data augmentation is to increase the amount of pertinent variance in a dataset. This is a challenging problem which must be approached carefully with prior knowledge of the source of the variance. Taking the example of lesion segmentation on brain images, additional examples of lesions could be produced through careful deformation of existing lesions. However, it would be important for such a procedure to follow specific rules which govern the appearance of real lesions. To create such a procedure which can be applied with confidence and generalises across all lesion examples would be difficult, time-consuming and highly application specific. One example of such a method is provided in [Shrivastava et al., 2016] for the purpose of gaze estimation. The authors simulated the pertinent variation by using a simulator which would generate synthetic, labelled, samples, which were then made to look realistic

through refinement. This was only possible as a complete model of the pertinent variance could be produced. The absence of such a model, as is the case in many medical imaging applications, makes such augmentation difficult and rare. Despite this, accurate augmentation of pertinent variance can be extremely valuable, as it directly reduces the need for additional training examples to be acquired. This is demonstrated by the substantial improvements seen in [Shrivastava et al., 2016].

As noted in [Krivov et al., 2017], there is a tendency within medical imaging to prefer the removal of non-pertinent variance as opposed to augmentation, and an acceptance that pertinent variance cannot be reliably augmented. This is partly due to the ease at which much of the non-pertinent variance can be removed, and partly due to the lack of suitable augmentation procedures for many of the sources of non-pertinent variance in medical images. This is reflected in [Kamnitsas et al., 2017b], where the authors choose to only employ reflection and intensity augmentation for brain lesion segmentation, with even the latter omitted when using larger datasets. On the other hand, in [Ronneberger et al., 2015], the authors strongly encourage the use of augmentation in their application of microscopy images, particularly the application of random elastic deformations. This demonstrates how careful consideration of the application will dictate which types of augmentation are appropriate. Random elastic deformations may be an appropriate model for microscopy images, in which the objects of interest (i.e. cells) are generally fluid and unconstrained. Applying the same procedure to brain images could lead to certain anatomical constraints such as symmetry, rigidity and structure being disregarded.

GANs offer a potentially valuable addition to the arsenal of augmentation techniques which are currently available. One of the main potential advantages of GANs is that they take many decisions away from the user, in much the same way as deep learning removed the need for "handcrafted" features. An ideal GAN will transform the discrete distribution of training samples into a continuous distribution, thereby simultaneously applying augmentation to each source of variance within the dataset. For example, given a sufficient number of training examples of images at different orientations, a GAN will be able to produce examples at any orientation, thereby replicating the effects of applying a rotation augmentation. While orientation is a source of variance which can easily be augmented or removed using traditional

methods, consider instead a more challenging source of variance such as ventricle size in brain imaging. Again, given a sufficient number of training examples of patients with different discrete ventricle sizes, a trained GAN will be able to produce examples along the continuum of all sizes, from the smallest to the largest in the dataset. To perform the same kind of augmentation using, for example, deformations would involve a complex model of realistic ventricle size and shape and impact on the surrounding anatomy. By simultaneously learning the distribution of all sources of variance, the GAN infers this model directly from the available data.

GAN augmentation should not, however, be considered a replacement for traditional augmentation, rather a method to augment sources of variance which are difficult to augment in other ways. One major advantage that traditional augmentation has over GAN augmentation is the ability to extrapolate, as well as interpolate. GANs can provide an effective way to fill in gaps in the discrete training distribution, but will not extend the distribution beyond the extremes of the training data. For example, the training data may have examples at orientations from $+90°$ to $-90°$, yet test cases could be at any orientation. In this case, traditional rotation augmentation would be necessary to extend the training data distribution to the other orientations. In general, appropriate traditional augmentation procedures should be used to extrapolate and extend the manifold of semantically viable images. GANs can then be used to interpolate between the discrete points on this manifold, providing an additional data-driven source of augmentation.

The main potential limitation of GAN augmentation is the ability of the GAN to generate images with a high enough image quality. While great improvements have been made in the field over the last few years, GANs cannot be relied upon to produce images with perfect fidelity. In [Chuquicusma et al., 2018], the authors demonstrate that their GAN generated lung nodules would routinely fool a radiologist with 4 years experience but would be frequently identified by one with 13 years experience. However, both [Dosovitskiy et al., 2015] and [Richter et al., 2016] demonstrate that perfect fidelity is not necessary to improve results with synthetic data. This is not a problem for traditional augmentation procedures which do not significantly degrade the images. Whether the advantage of additional data is outweighed by the disadvantage of poorer quality images is one of the questions we hope to address in this chapter.

5.3 Contribution

The results reported in [Amitai and Goldberger, 2018, Moradi et al., 2018] suggest that GANs can have a significant benefit when used for data augmentation, though this is still a relatively unexplored area. In this chapter, we investigate this use of GANs more thoroughly for the purpose of medical image segmentation. An in-depth investigation into the effects of GAN augmentation is first carried out on a complex multi-class Computed Tomography (CT) Cerebrospinal Fluid (CSF) segmentation task using two segmentation architectures. By choosing not to co-register the images in this dataset, we are able to examine how GAN augmentation compares and interacts with rotation augmentation.

The transferability of the method is then evaluated by applying it to a second dataset of Fluid-attenuated Inversion Recovery (FLAIR) Magnetic Resonance (MR) images for the purpose of single-class White Matter Hyperintensity (WMH) [Wardlaw et al., 2013] segmentation. This is a well-studied problem and poses challenges typical to medical image segmentation tasks. WMH can be split into White Matter Hyperintensity of Presumed Vascular Origin (WMH_{pvo}) [Wardlaw et al., 2013], and stroke lesions, though for this application, they are treated as a single class.

Aside from establishing whether GAN augmentation can lead to an improvement in network performance, we answer the following 5 important questions:

- Does the choice of segmentation network architecture change the presence or degree of improvement?
- How does GAN augmentation compare to traditional augmentation methods?
- Does the amount of synthetic data added affect this improvement?
- Does the amount of available real data affect this improvement?
- Does the approach generalise to multiple datasets?

We also explore the distribution of generated images to better understand what modes of augmentation are provided. This allows us to confirm that the GANs are producing images which are different from those in the dataset. We also show how the procedure can generate images with the same pathology, but different unrelated anatomy, and vice versa, demonstrating the ability to perform these particularly challenging forms of augmentation. Finally, we examine whether GAN augmentation leads to novel images unseen in the dataset.

5.4 Methods

We use a Progressive Growing of GANs (PGGAN) network [Karras et al., 2017] to generate synthetic data. PGGAN was chosen on the basis of its training stability at large image sizes and apparent robustness to hyperparameter selection, with only minor changes to the hyperparameters suggested in [Karras et al., 2017] proving suitable for all experiments. Whether the choice of GAN architecture will affect the quality of the augmentation is unclear, however there is evidence [Lucic et al., 2017] to suggest that different GAN architectures produce results which are, on average, not significantly different from each other under optimal hyperparameter selection, with the main difference between the methods being the ease at which these hyperparameters can be found.

We train a PGGAN on 80k patches sampled from the available training data as a preprocessing step prior to training a segmentation Convolutional Neural Network (CNN). The PGGAN is trained on multi-channel image patches containing both the acquired image and manual segmentation label, thereby learning the manifold containing this joint data. Synthetic examples are then sampled randomly from this manifold using the trained generator and used to augment the same 80k patches upon which the GAN was trained, forming the training data used when training the subsequent segmentation network. The standard generator architecture is modified to output an image with the required channels (image + one or more segmentation channels). The discriminator architecture is also modified to accept such multi-channel images as input. The segmentation channels, whilst binary in nature, are mapped onto the same con-

tinuous range as the image channels (i.e. mapped from $\mathbb{Z}_2$ to $\mathbb{R}$) to allow both the image and segmentation channels to be processed within the same convolution operator. The only other alterations to the default PGGAN parameters are to increase the number of images shown at each resolution level from 600k to 800k, and for CT experiments, to concatenate a 32x32 layer of Gaussian noise at the start of the fourth (32x32) resolution level. This change was found to aid in producing CT images with a more realistic noise pattern.

Each segmentation network was evaluated using a training, validation and test set, with the training set consisting of the 80k patches sampled from the available training images with any additional synthetic data added on top of this. During training, performance (as measured by Dice Similarity Coefficient (DSC)) on the validation set was monitored, with the best model at the conclusion of training applied to the test set.

5.5 Experiments

A set of experiments were designed to assess the effect of introducing GAN derived synthetic data to a segmentation task. In these experiments, a number of key variables were modified:

- Amount of available real data: To simulate a situation with limited training data, the amount of real data was artificially reduced by randomly selecting a percentage of the available training data, prior to sampling the 80k training patches. We explored both a moderate (50%) and extreme (90%) reduction in available data. Note that this reduction in available data is enforced for both the GAN and segmentation network training stages, ensuring the GAN is never exposed to more labelled data than the corresponding segmentation network.

- Amount of additional synthetic data: To investigate whether the amount of synthetic data added to the real data affects the performance of a segmentation network, experiments

were run with different amounts of additional synthetic data. To ensure equal access to the information available in the real data between experiments, synthetic data is added to the real data, increasing the size of the dataset, rather than replacing real data. The amount of additional patches is therefore expressed as a percentage of the real patches. For example an experiment with +50% synthetic data would use 120k patches (80k real and 40k synthetic).

- Dataset: Two different datasets are explored to assess the ability for GAN augmentation to generalise across segmentation tasks. The first dataset contains CT images with manually delineated CSF labels split into into 3 classes: cortical CSF, brain stem CSF and ventricular CSF. Data is split in the same way as in [Chen et al., 2018a], using the same preprocessing and sampling procedures. This provides 500 manually labelled training images, with an additional 282 validation images, from 101 subjects. For these experiments, the average DSC is used as the primary measure of performance, though results across each class are also analysed.

  The second dataset contains MR FLAIR images with manual binary WMH segmentations. 147 FLAIR images were acquired as described in [Valdes Hernandez et al., 2015] (see Section 2.6.2 for further details). These were manually segmented, before being bias corrected, brain extracted, rigidly co-registered and intensity normalised (as described in Section 4.3.2), and randomly split into equal sized training, validation and test sets.

  By selecting two dissimilar tasks (multi- and single-class segmentation) across two modalities (CT and MR) we cover a wide range of likely applications for GAN augmentation.

- Segmentation network: We investigate three different segmentation networks across the experiments. In [Chen et al., 2018a], the authors show that both UNet and Residual UNet (UResNet) [Guerrero et al., 2018] architectures perform well on this CT dataset, we, therefore, choose to explore both of these. The same hyperparameters were used as in [Chen et al., 2018a]. DeepMedic [Kamnitsas et al., 2017b] is a popular general-purpose segmentation algorithm which has been shown to perform well in many applications and was therefore chosen as a third network to explore. DeepMedic was modified only so as

to accept 128x128 2-dimensional (2D) patches.

Between these three, we represent the most popular CNN architectures currently in use.

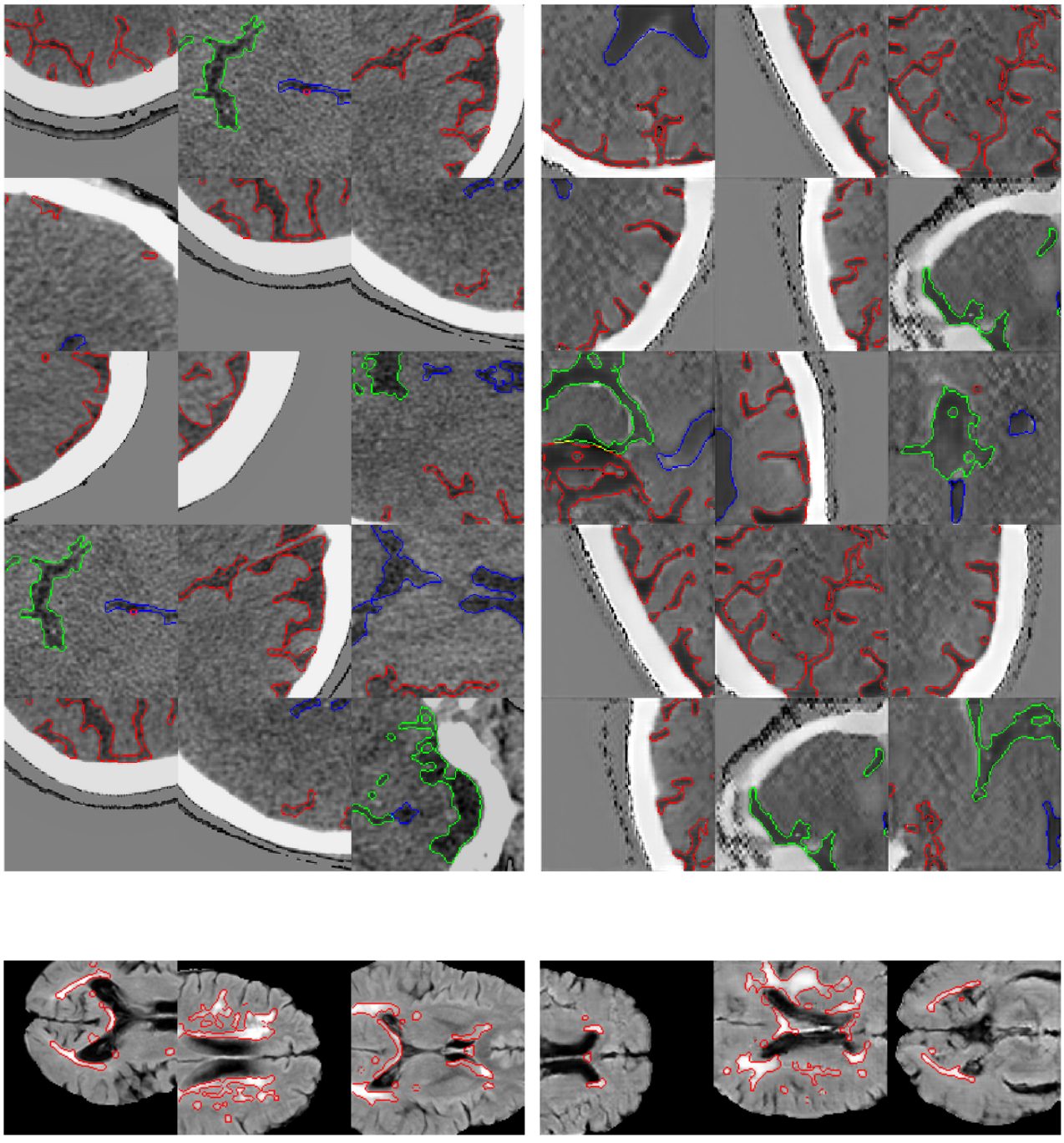
- Augmentation: As discussed in Section [Krivov et al., 2017] extensive augmentation, beyond simple reflection, is rarely used in brain imaging due to the variety of preprocessing options available and anatomical constraints of the brain. However, in order to examine the interaction of GAN and rotation augmentation we elect not to perform coregistration on the CT dataset. Of the other common forms of augmentation, reflection augmentation is routinely performed in all experiments, translation augmentation is encapsulated in the patch based approach, intensity augmentation is obviated by normalisation, and deformations are not considered due to anatomical constraints (preserving shape, symmetry etc.).

Table 5.1: Summary of experiments

% available real data	% added synthetic data	Segmentation network	Dataset	Augmentation type	Repetitions
100, 50, 10	0, 50, 100	UNet, UResNet	CT	Rotation+GAN	8
100, 50, 10	0, 100	UNet	CT	None, GAN, rotation, rotation+GAN	8
100, 50, 10	0, 12.5, 25, 37.5, 50, 100	UNet	CT	Rotation+GAN	8
100, 90, 80, 70, 60, 50, 40, 30, 20, 10	0, 50	UNet	CT	Rotation+GAN	8
100, 50, 10	0, 50, 100	DeepMedic	MR	GAN	14

Table 5.1 summarises the 5 sets of experiments which were carried out to answer the questions posed earlier. In each experiment, the segmentation network is treated as a black box and unchanged. This provides a fair platform upon which to observe the effects of GAN augmentation by ensuring that any changes in performance are as a result of the additional synthetic data, and not of changes in the network itself. The same PGGAN architecture is used in each experiment, configured to produce images with a size of 128-by-128px. GAN training took 36 hours, each UNet took 4 hours, each Res-UNet took 24 hours and each DeepMedic network

took 24 hours on an Nvidia GTX 1080 Ti or similar GPU. Examples of real and synthetic patches generated for each dataset can be seen in Figure 5.1.



(a) Real (b) Synthetic

Figure 5.1: Examples of real and GAN generated synthetic patches for each dataset. *Top:* CSF. Red: Cortical CSF. Green: Brain stem CSF. Blue: Ventricular CSF. *Bottom:* WMH.

5.6 Results

5.6.1 Segmentation results

The following tables and graphs show the results over the two sets of experiments. All tables show the average DSC, with the standard deviation in brackets. Results which are statistically different (2-tailed t-test, 5% significance level) from those observed when no additional data is introduced are shown in bold.

Table 5.2: **CSF segmentation:** Results with different proportions of the available training data and varying amounts of additional synthetic data using UNet and UResNet architectures.

		Available data					
		UNet			UResNet		
		100%	50%	10%	100%	50%	10%
Additional Data	0%	88.9 (0.51)	86.0 (0.50)	76.9 (0.58)	86.8 (0.82)	82.7 (1.55)	72.5 (1.98)
	50%	89.2 (0.30)	**87.3** (0.46)	**78.6** (1.04)	86.3 (1.44)	**84.3** (1.31)	74.3 (1.63)
	100%	89.3 (0.39)	**86.9** (0.36)	**78.4** (0.99)	86.3 (1.24)	84.1 (1.32)	**74.7** (1.18)

Table 5.3: **CSF segmentation:** UNet results with different proportions of the available training data and different augmentation techniques.

	Available data		
	100%	50%	10%
No augmentation	88.1 (0.32)	85.0 (0.58)	75.1 (0.60)
GAN augmentation	88.4 (0.41)	85.6 (1.33)	76.3 (1.77)
Rotation augmentation	**88.9** (0.51)	**86.0** (0.50)	**76.9** (0.58)
GAN + Rotation augmentation	**89.3** (0.39)	**86.9** (0.36)	**78.4** (0.99)

Table 5.4: **WMH segmentation:** Results with different proportions of the available training data and varying amounts of additional synthetic data.

		Available data		
		100%	50%	10%
Additional Data	0%	66.0 (1.26)	61.4 (2.67)	52.2 (6.65)
	50%	65.5 (1.21)	**63.7 (0.69)**	**57.2 (4.09)**
	100%	**64.8 (1.34)**	62.8 (1.17)	55.7 (4.26)

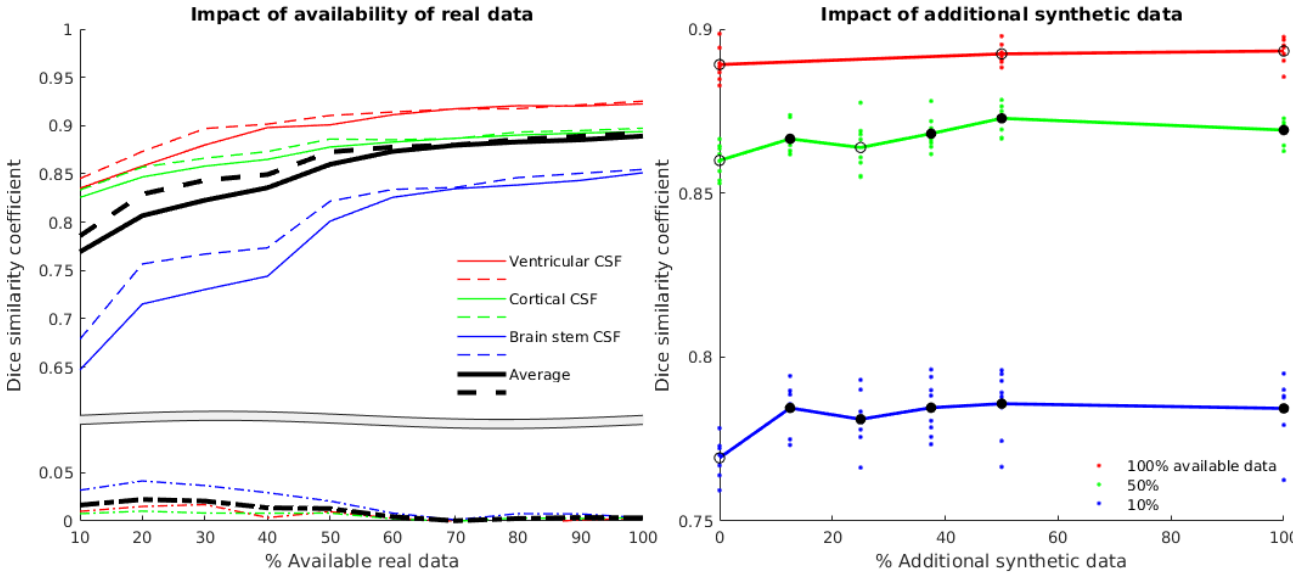


Figure 5.2: **CSF segmentation:** *Left:* Average DSC for each class (coloured) and mean across classes (black) as availability of real data varies. Solid lines show performance without GAN augmentation, dashed lines show performance with +50% synthetic data, and dot/dashed lines show the improvement seen with GAN augmentation. *Right:* Average DSC observed using a UNet as synthetic data is added, when 100%, 50% and 10% of the total amount of real data is used. Each coloured dot represents an experiment. Black circles show the mean with filled circles indicating results significantly different from those without any additional synthetic data as found through a 2-tailed t-test with a significance level set at $p < 0.05$.

5.6.2 Qualitative evaluation

Whilst various metrics exist for quantitatively evaluating synthetic images directly, these would have little meaning in this case, especially as we have already indirectly evaluated the images

through their application in data augmentation. Therefore, in addition to the quantitative segmentation results, the generated images MR were also examined qualitatively to gain an understanding of what extra information is being provided through GAN augmentation. Figures 5.3, 5.4 and 5.5 show samples of generated images (left of pair) along with their nearest neighbours in the training dataset (right of pair) for the GANs trained on patches from 5, 25 and 50 images respectively. These images were examined by the author looking for three features: Cases where lesions were duplicated on different anatomy; cases where lesions were changed whilst anatomy stays the same; and cases where the nearest neighbour in the dataset is substantially different from the synthetic image. The final feature is a sign that the GAN has learned a manifold which contains some smooth regions which enables interpolation between certain images, potentially including novel anatomy. A selection of these cases are indicated using green arrows (same lesions, different anatomy), yellow arrows (same anatomy, different lesions), and blue dots (completely new anatomy and lesions).

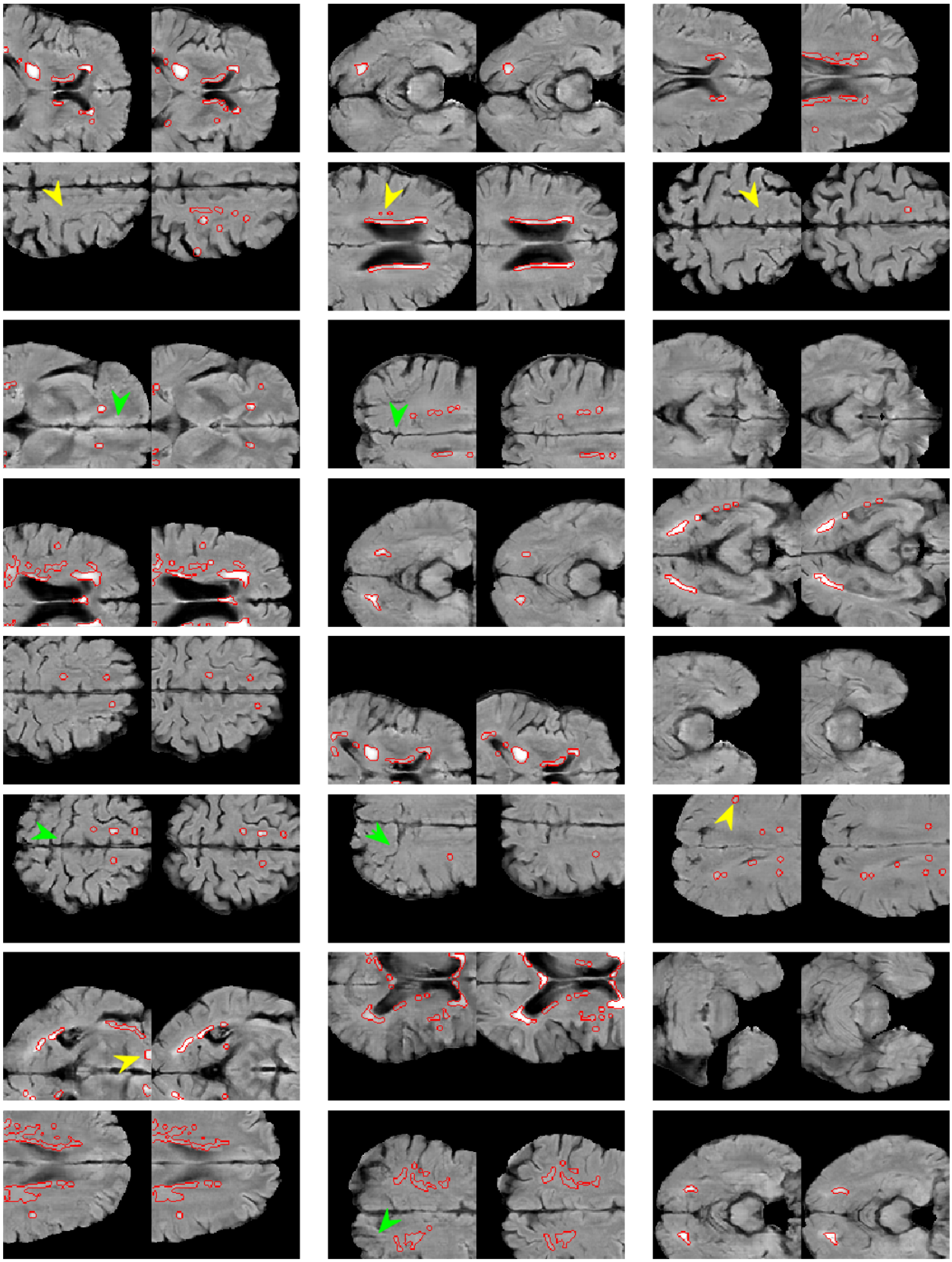


Figure 5.3: **5 training images**

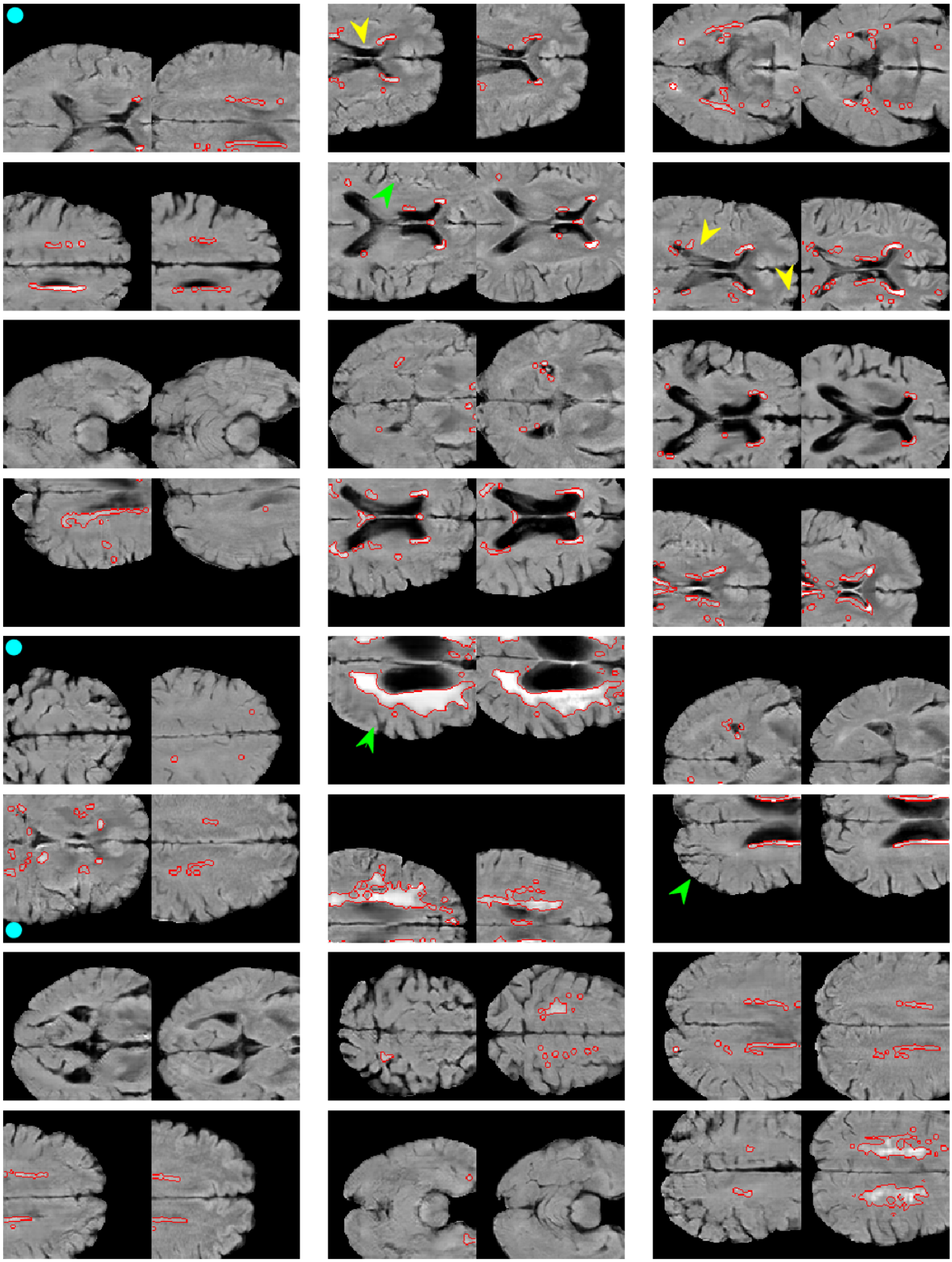


Figure 5.4: **25 training images**

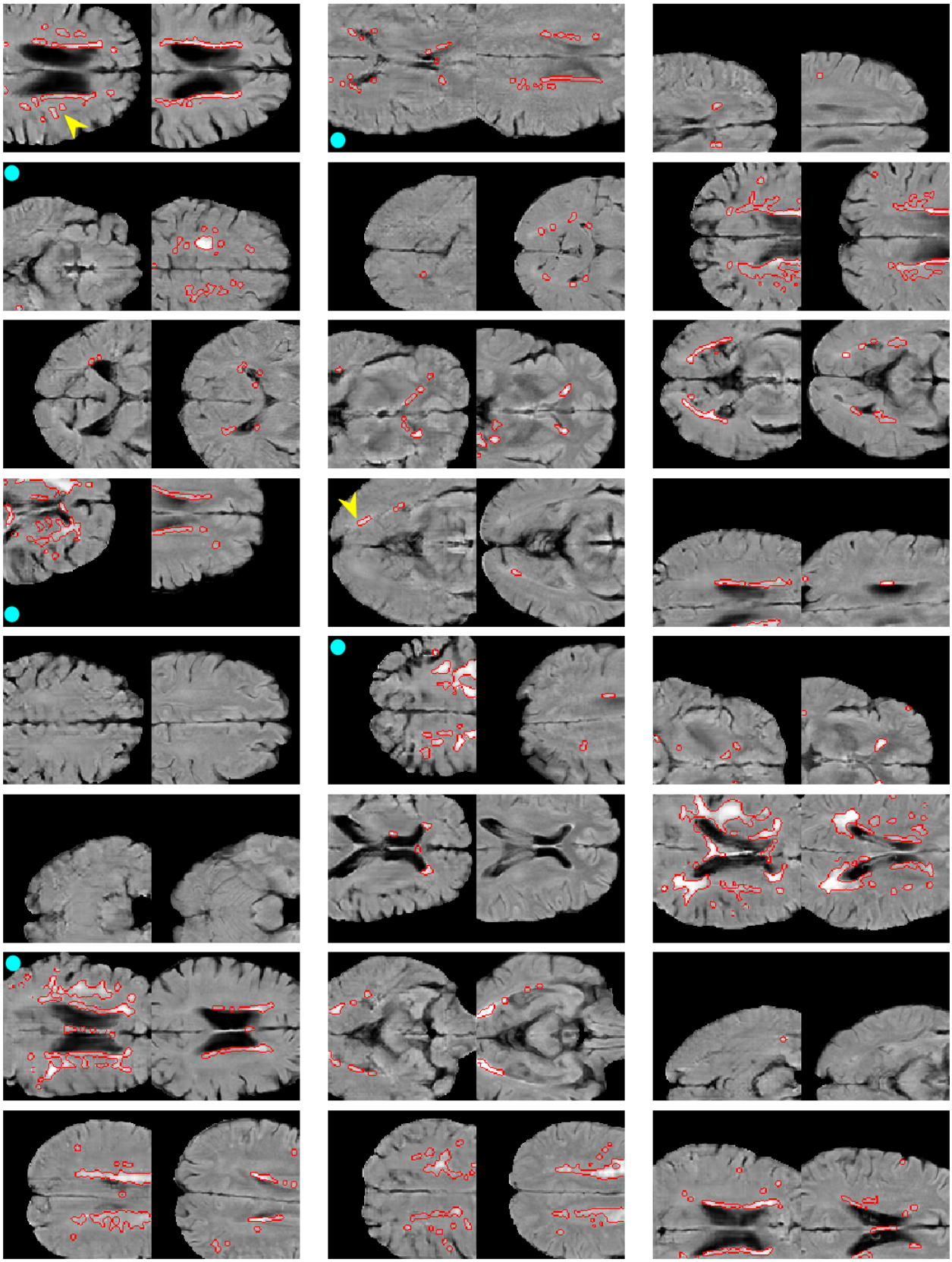


Figure 5.5: **50 training images**

5.7 Discussion

It can be seen from across all of the results that GAN augmentation can provide a modest but significant improvement in segmentation performance (when measured in terms of DSC) in many cases. By far the strongest factor controlling the improvement seen is the amount of real data available for training. There is a clear trend across all results that the greatest improvements can be seen in cases where real data is limited. However, Figure 5.2 suggests that there is perhaps a drop in improvement seen at the very lowest levels of available data. This is possibly a result of there being too little data to properly train the GAN. Results on the CSF data suggest that there are no circumstances in which using synthetic data leads to worse results even when large amounts of real data is available. However, this is not reflected in the WMH results in Table 5.4, where a loss in DSC is observed when all available data is used. This suggests that there may be a tipping point associated with the amount of available data, beyond which GAN augmentation harms rather than helps. This indicates that beyond this point the synthetic data is not as useful as the data from which it is derived, implying that the GAN does not completely reproduce the information from the training data. This could either be as a result of lower image quality, or reduced variation. Care must, therefore, be taken when applying this method in cases where plenty of real data is available. The exact point at which synthetic data is no longer useful is likely application specific, and could only be calculated through a series of experiments, gradually introducing additional real data. Such an approach would be time-consuming as a new GAN would be required for each level of data. However, for practical applications, it would not be necessary to find the exact tipping point, since it is only important to know if GAN augmentation provides a benefit or not when all available data is used.

The added benefit of GAN augmentation in cases of limited data can also be seen in the DSC observed on the individual CSF classes in Figure 5.2. A ratio of 1.35:4.35:1 between ventricular, cortical and brain stem CSF classes in the training set indicates a moderate class imbalance, with examples of the former and latter being relatively limited. Figure 5.2 shows that it is these two classes which benefit most from GAN augmentation. Of these, brain stem

CSF segmentation appears to benefit the most, though this can be attributed to ventricular CSF segmentation being an inherently easier proposition, and therefore being consistently well segmented anyway.

Table 5.2 shows that there is little difference in the effect of GAN augmentation when using different CNNs. This, coupled with the WMH results, provides evidence that GAN augmentation can benefit any segmentation network, regardless of architecture.

Figure 5.2 suggests that the amount of additional synthetic data makes little difference to any improvement seen. There are no significant differences between any pair of results using different quantities of synthetic data across all experiments. The observation that the approach is robust to the amount of synthetic data added is welcome, as it suggests that the amount of synthetic data is not an additional parameter which needs to be finely tuned. This, coupled with the earlier observation that synthetic data does not impair performance when real data is limited, makes GAN augmentation a practical proposition.

It is interesting to note that the improvements returned by using both traditional and GAN augmentation, as seen in Table 5.3, are consistently more than the sum of the improvements given by using the two methods separately. Whilst these differences are individually not significant, this result provides strong evidence that the additional information provided by the two augmentation methods are independent. It also suggests that when used together they are potentially synergistic, an observation which agrees with the results in [Amitai and Goldberger, 2018]. This supports the hypothesis that GANs provide an effective alternative to traditional augmentation when attempting to interpolate within the training distribution, but cannot extrapolate beyond its extremes without the aid of traditional augmentation like rotation.

Figures 5.3, 5.4 and 5.5 provide an interesting insight into what additional information is being provided by GAN augmentation. In the case of 5 training images (Figure 5.4), it is clear that each generated image is based heavily on an image from the training set. This is not surprising since there are very few images to train on, and little variation which can be learned. However, there are subtle differences present in the majority of synthetic images, either in anatomy or in pathology. There are cases where lesions present in the real image are not reproduced in the

synthetic image, as well as cases where the shape and number of lesions present in the synthetic image differ from those in the real image. Both of these effects can be extremely valuable to prevent overfitting when training a model - the former decoupling the presence of lesions from the surrounding anatomy, and the latter providing more variety of pathology. In Figure 5.4, we can see that when the number of training images increases to 25, we begin to see cases where there are no close matches in the training set, in addition to the cases of novel anatomy and pathology seen previously. This trend gets even stronger in Figure 5.5 where all 50 training images are used. There are often substantial differences between the synthetic images and their closest real image, suggesting that the GAN has learned to produce data substantially beyond what was provided to it. It is also reassuring to observe that these modifications appear reasonable in all cases, with no obvious unrealistic lesions or anatomy being synthesised.

5.7.1 Conclusion

This chapter has presented a method for augmenting training data using GAN derived synthetic images and demonstrated that this can improve results across two segmentation tasks. The method has been shown to work best in cases of limited data, either through a lack of data or as a result of class imbalance. Further experiments are required to fully assess its suitability in other domains, though it has the potential to be a practical preprocessing step in a wide range of applications. Applying GAN augmentation requires little overhead, involving only the training of a single out-of-the-box GAN, does not involve optimising additional parameters and has been shown to be low-risk by never damaging performance when training data is limited. The exact amount of improvement expected is likely a complex function of the amount of real data available, the amount and quality of synthetic data and the task itself. Many more experiments are required to fully understand this relationship, however a conservative interpretation of the results from the two typical segmentation tasks explored here suggests that in cases where $5 - 50$ labelled images are available, augmenting patches sampled from these with an additional $10 - 100\%$ GAN derived synthetic data has the potential to lead to significant improvements in segmentation results as measured by DSC. Future work will involve

investigating this relationship further, as well as in other areas such as classification, and to evaluate the impact of different GAN architectures and parameters.

The approach taken in this chapter is similar to that of traditional data augmentation - no additional information is added beyond that which was already present in the dataset. However, the use of GANs offers the potential to incorporate more information into this process. The next chapter looks to investigate this further by modifying the training procedure of a GAN to allow for unlabelled data to be used in addition to the available labelled data, with the aim of further improving the performance of supervised segmentation algorithms.

Chapter 6

GANsfer Learning: Combining labelled and unlabelled data for GAN based data augmentation

6.1 Introduction

Having previously demonstrated that Generative Adversarial Networks (GANs) can be used to produce labelled images of high enough quality to perform learning on, and that these images can be used for effective data augmentation in Chapter 5, we now consider whether this approach can be improved by introducing unlabelled data. One of the limitations of training GANs on relatively small amounts of data is that the resulting learned manifold will be characterised by a small number of modes. The generator will, therefore, produce images from around these modes, with only subtle variations. These subtle changes have proved sufficient to reduce overfitting when performing data augmentation in Chapter 5, and in the other applications reviewed in Section 3.3. However, there is clearly scope to improve upon this. To be able to learn a smooth manifold, allowing for interpolations between points describing real images, a critical mass of data is required. However, if such a large amount of labelled data is available, there is likely little need to perform data augmentation and training a GAN

becomes redundant. Instead, we wish to move towards learning this smooth manifold from a substantially smaller number of labelled images. We propose to do this by leveraging a large amount of unlabelled data in addition to the limited labelled images, using a technique inspired by neural network transfer learning.

Transfer learning has proved to be an effective approach to neural network training across many applications, including medical imaging [Shin et al., 2016]. It involves training a model on a separate, usually very large, dataset, then applying it to the desired problem for which there is comparatively little available data. This is usually done with some form of fine tuning, where all or part of the pre-trained network is subsequently trained on the samples from the target problem. The intuition behind this is that many of the useful low-level features will be shared across tasks. At the lowest levels, these might be looking for simple edges and colours, while slightly higher level features might be looking for line segments. It is only deeper in the network that these features take on any meaning relating to the actual domain and, even deeper, specific task. It, therefore, seems sensible to learn these low-level features on a separate dataset. As demonstrated in [Shin et al., 2016], this dataset need not even be related to the original dataset. This has several advantages including reducing computation time due to model reuse, and increasing the quality of the learned features, especially in cases where there may not be sufficient data to otherwise learn these optimal low-level features.

This process can be thought of as learning (or reusing) a general purpose feature extractor, followed by learning how to interpret these features in the context of the task at hand. Feature extraction and interpretation are decoupled and learned separately. We propose a similar framework for training a GAN where we aim to decouple the learning of anatomical variation and the learning of appearance. In the case of labelled images, appearance encompasses not only pixel intensities in image space, but also the corresponding binary segmentation labels contained in additional channels. The aim is therefore to learn the parts of the network which correspond to the appearance from a small amount of available labelled images, while the parts responsible for generating realistic anatomy are trained on a larger amount of unlabelled images.

Recent work [Madani et al., 2018] showed that incorporating unlabelled data using a GAN

framework can lead to significant improvements in accuracy in a chest X-ray classification task. The authors repurpose the discriminator as a classifier, forcing it to output a belief as to whether an input image is diseased or not, as well as whether it is real or synthetic. In this way, the discriminator uses unlabelled data to improve its ability to identify synthetic images, thereby learning additional features which are useful for its secondary role of distinguishing healthy from diseased images. By training on both tasks concurrently, the discriminator must remain good at both tasks. A somewhat similar approach was presented in [Ross et al., 2018] for endoscopic video segmentation. A conditional GAN using a UNet-like generator is first trained on the unlabelled data to perform an arbitrary auxiliary task - in this case, an image re-colourisation procedure. The goal of this is to pre-train the first half of the UNet as an effective feature extractor by leveraging the unlabelled data. The labelled data is then used to fine-tune the UNet in the target segmentation task. An increase in classification accuracy from 51% to 73-76% when labelled data is limited in [Madani et al., 2018], and in Dice Similarity Coefficient (DSC) from 0.57 to 0.65 in [Ross et al., 2018] shows the potential for incorporating unlabelled data into a classification system.

6.2 Methods

6.2.1 Rationale

We must first consider the structure of a GAN, the generator in particular, to understand which layers are responsible for generating the anatomical structures, and which are responsible for producing the realistic appearance of these structures. For the purposes of these experiments we use the Progressive Growing of GANs (PGGAN) [Karras et al., 2017] architecture, which aims to build a typical Deep Convolutional Generative Adversarial Network (DCGAN)-like architecture gradually by adding additional layers during training. At the conclusion of training, a possible generator architecture for a small network will look like that shown in Figure 6.1. If we consider the final convolution process, Figure 6.2, we see that each of the output channels is simply a weighted sum of the feature maps from the penultimate layer. Therefore,

when generating multi-channel output, with an image and one or more separate corresponding segmentation maps (binary or continuous), we can observe that the features are semantically linked to the structures present in the image (Figures 6.3 and 6.4).

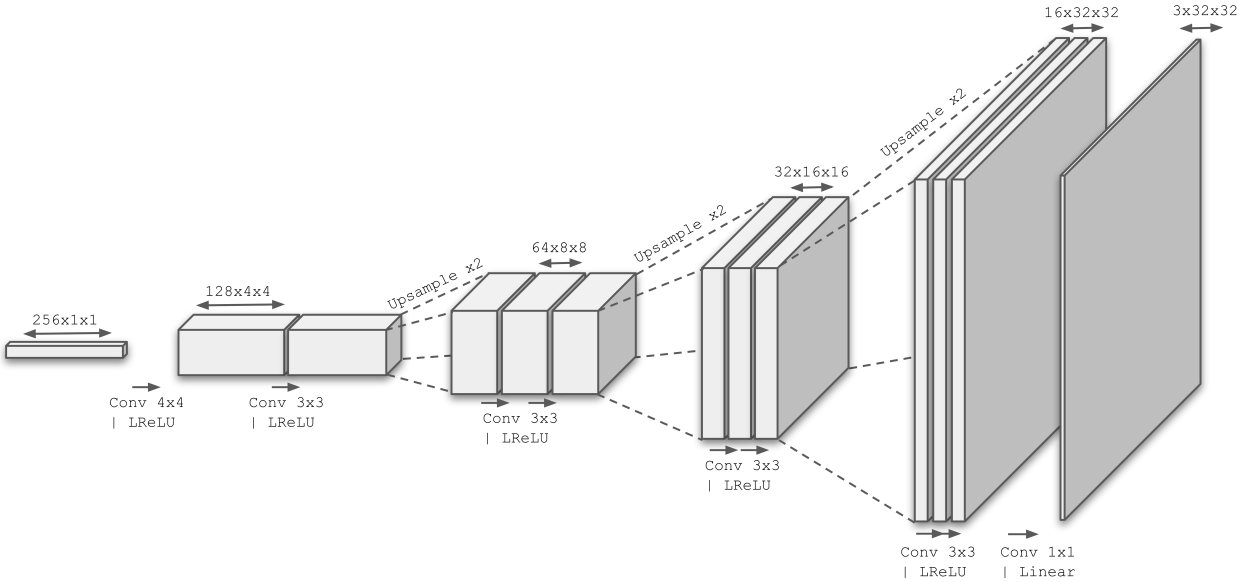


Figure 6.1: Architecture of a typical PGGAN generator for a 3 channel 32-by-32px image from a 256 element latent vector.

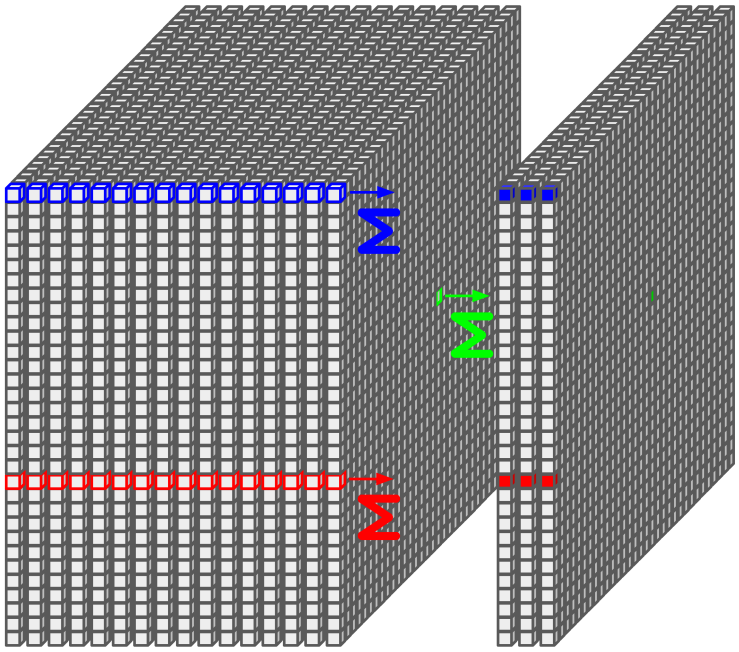


Figure 6.2: Final layer of the architecture from Figure 6.1 in greater detail showing how each channel in the final image is a weighted sum of the elements from the penultimate layers.

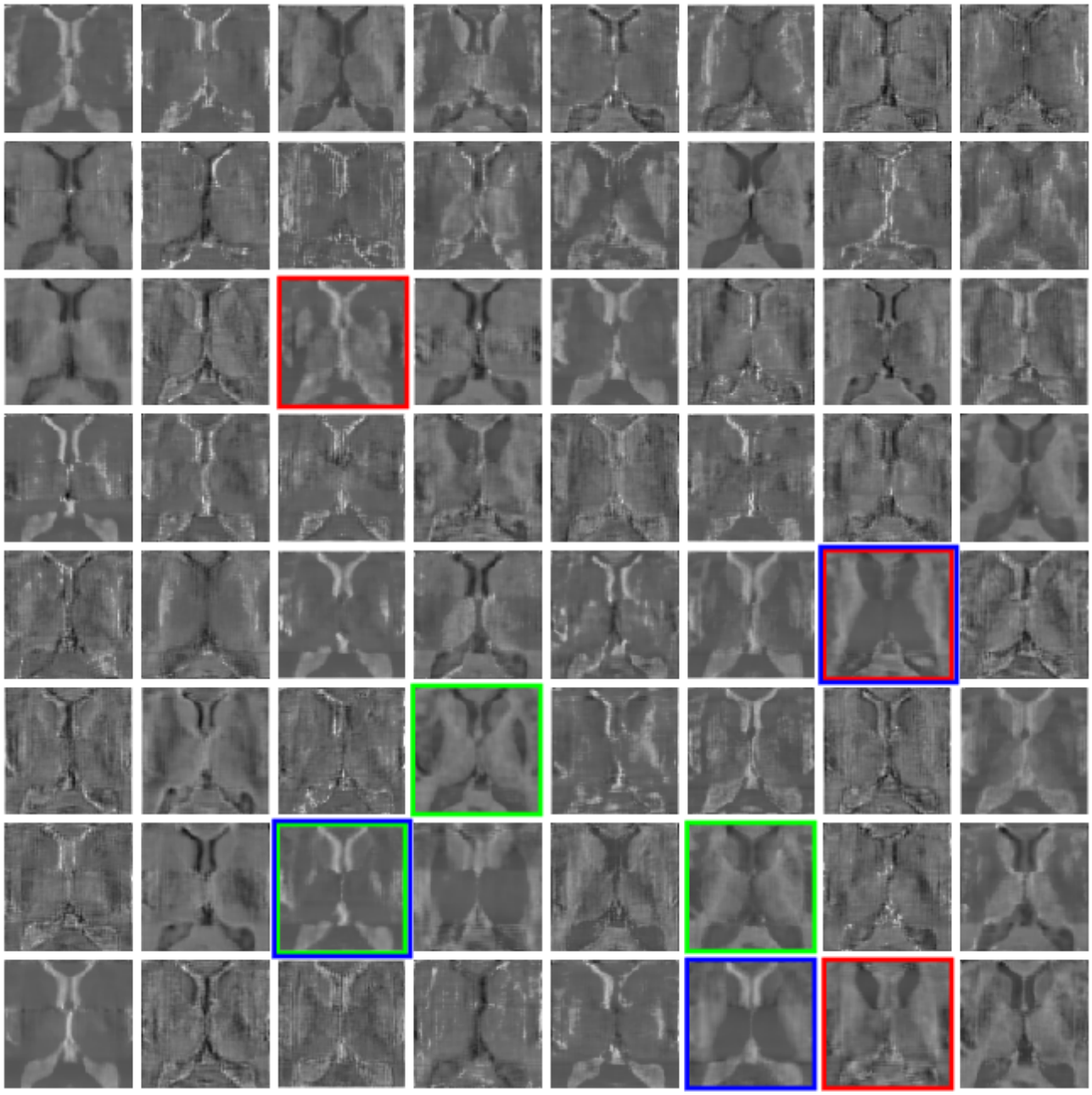


Figure 6.3: Penultimate layer feature maps generating a Magnetic Resonance (MR) image patch. A linear combination of these patches is used to generate the final image. Note that some feature maps have particularly high contrasts between certain structures, indicating their use in producing these structures in the image and segmentation channels. Construction of the final MR image and segmentation channels from these maps is shown in Figure 6.4. The three maps with the strongest absolute corresponding weight for each visible segmentation channel are shown. Red: Caudate. Blue: Thalamus. Green: Putamen. Note that some maps contribute to multiple segmentation channels.

This is a very useful representation. The final layers generate a set of feature maps which can be thought of as a set of image-specific bases which can be combined in different pre-learned ways to yield both the image and corresponding segmentation channels. This shows that the

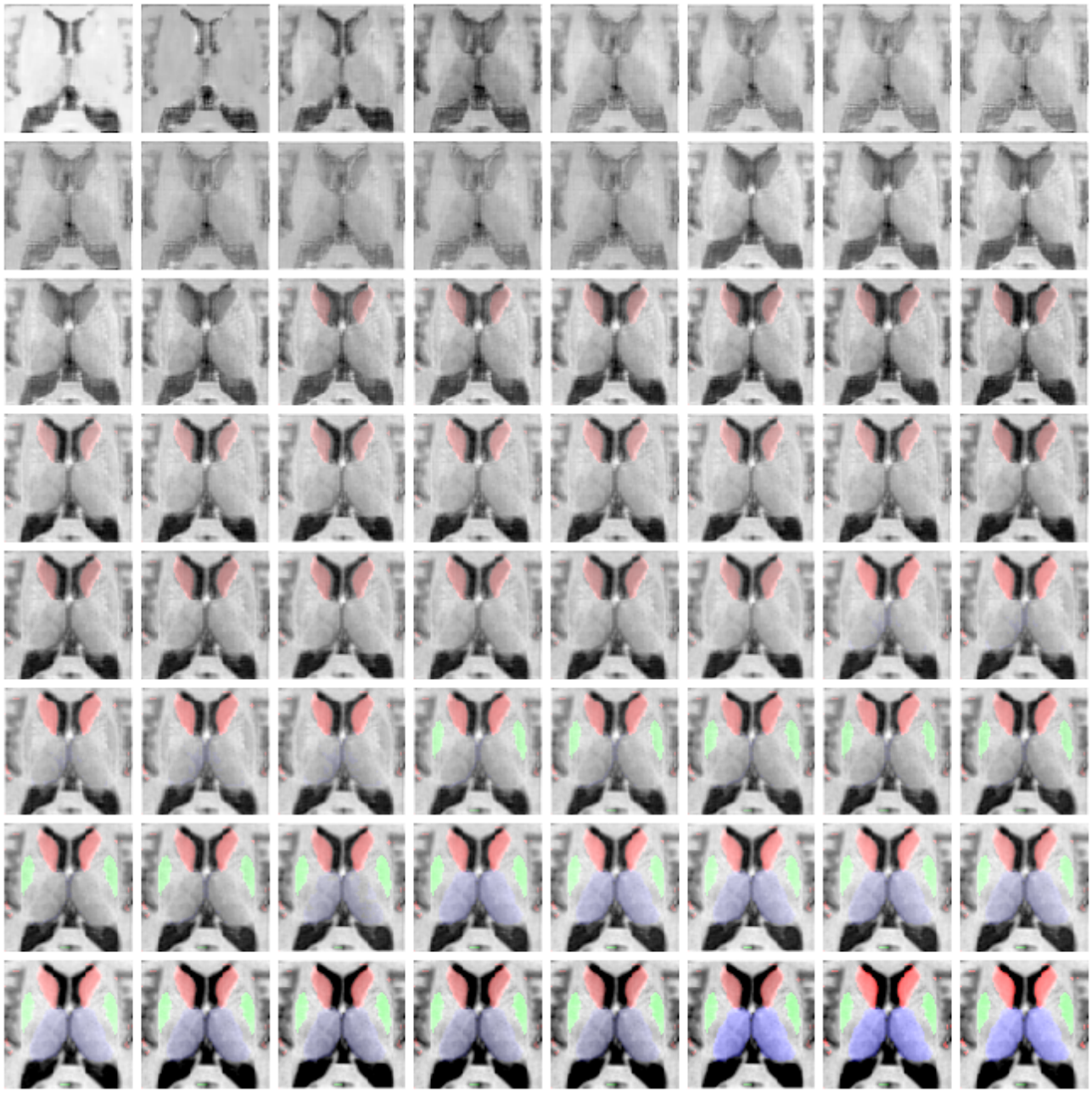


Figure 6.4: The process of constructing an MR image and segmentation channels using a linear combination of the feature maps shown in Figure 6.3. Read top-to-bottom, right-to-left the output image using only feature maps up to that point are shown. Note how structures and segmentations are introduced individually. MR and segmentation channels are scaled at each stage for visualisation.

segmentation maps are only generated in the final layers, and they are generated from the same feature maps which produce the image. This also follows when we consider that the PGGAN grows over time during training, with the earlier layers tasked with generating low-resolution images, and therefore the low-frequency information such as the anatomy, and the later layers responsible for the higher frequency information such as texture. This can also be thought of

intuitively by considering Figure 6.3: If it is possible to generate a set of feature maps which can be combined to produce different anatomy, it follows these will produce sensible segmentations when combined in the same way as in Figure 6.4.

Having established that the final layers are responsible for the appearance, and therefore the segmentation maps, and that the earlier layers are responsible for the anatomical variance, we can consider how to train these components separately. A traditional transfer learning approach would involve training on the large unlabelled dataset first, and then refine the final layers on the smaller labelled dataset. However, this would involve increasing the number of output channels at the refinement stage, not only changing the desired output distribution but also its dimensionality. This sudden change in the discriminator objective function was found to confuse the network and cause a failure to converge. This is because the information required to form the segmentations may not be present in the final layers, leading the network to have to either effectively turn the final layers into a very small segmentation network, or to restructure the entire network to allow for this information to be passed forward from the earlier layers where the image anatomy is defined. We, therefore, propose to pre-train the GAN using the small labelled dataset, and refine the early layers using the large unlabelled dataset, before fine-tuning using a combination of both datasets. This avoids radically changing desired output distribution, allowing for a smoother transition, and pre-conditions the network to ensure sufficient information is present in the final layers to form the segmentation maps. The details of this process are given below.

6.2.2 GANsfer Learning

Phase 1: Train the GAN until convergence using the labelled data.

The goal of this phase is to pre-train the generator and condition it to ensure the necessary information to generate segmentations is present in the final layers. After this step, image diversity will be low due to limited training data, however, image and segmentation quality will be high.

Phase 2: Fix the weights of the final three resolution levels of the generator and train a new discriminator using the unlabelled data and the image channel from the generator output.

The goal of this phase is to learn more anatomical variation from the unlabelled data. Having learned a set of weights in the final layers which produce segmentation maps, we can increase image diversity by freezing these layers and continuing to train on the unlabelled data. Since the real data are now only single-channel images, and the trained discriminator expects a multi-channel image, we replace this with a new discriminator. This discriminator takes as input the training images and the image channel from the generator output. To allow this new discriminator to catch up with the generator, regular GAN training is performed using a ratio of 100 discriminator updates to 1 generator update for the first 5 training cycles, mirroring the pattern proposed for the start of training in [Arjovsky et al., 2017].

Phase 3: Train another discriminator using an equal mix of the segmentation channels of labelled images and of the generated images after phase 2 as ground truth, and the segmentation channels of the generator. Add this as a second discriminator to the GAN and perform regular GAN training by training both discriminators and updating the generator according to the loss from both discriminators. Gradually unfreeze layers of the generator up to the final layer.

The goal of this phase is to improve the quality of the generated images. The previous phase taught the generator to produce greater anatomical variation, but this comes at a cost of image quality due to the frozen layers. In this phase the frozen layers are gradually unfrozen, allowing these weights to be re-optimised with respect to the earlier layers. This is similar to the original training procedure of the PGGAN where layers for increasing resolutions are added in turn. This phase can, therefore, be thought of as a second pass through the GAN, optimising each layer in turn, thereby smoothly reattaching the newly trained early layers to the final layers. The final convolution layer remains frozen to ensure the image and segmentation channels remain coupled. During this phase the GAN is trained using two discriminators. The first is retained from *phase 2* and ensures the generator retains its ability to produce varied images. The second is a newly trained network which is trained purely on segmentation channels and ensures the quality of the segmentation channels is preserved. Its training set consists of the segmentation channels

from the labelled dataset, combined in equal parts with the segmentation channels from a new dataset which is generated using the generator from the end of *phase 2*. This "self-teaching" procedure ensures that the generator is encouraged not to forget the variation it has learned previously. Without this, there would be a significant disparity in the anatomical variation contained within the image dataset and segmentation dataset. Under these circumstances, the two discriminators would be in conflict, with the image based discriminator encouraging greater variation, and the segmentation based discriminator encouraging reduced variation.

In summary, *phase 1* ensures the final layers of the generator produce feature maps which can be linearly combined into images and segmentation maps. *Phase 2* trains the early layers to generate increased anatomical variability. *Phase 3* refines the generator to improve image quality. The data produced by the generator can then be used for data augmentation. This whole process is demonstrated in Figure 6.5. All network architectures are as described in Section 3.1.1, configured to operate on 80-by-80px images with the appropriate number of channels, and are trained using default hyperparameters and mirror data augmentation. Training is performed in *phase 1* for 360k images per resolution level, with another 360k for each transition period between levels (3600k total). *Phase 2* training is then performed for 120k images, with the last 3 generator up-sampling blocks and final layer frozen. Finally, *phase 3* training is performed for 180k images, with the third from last up-sampling blocks unfrozen for the first 60k, the second from last unfrozen for the next 60k, and all parameters except the final layer unfrozen for the final 60k. Training times were set to be the minimum time necessary to achieve the goals of each stage.

6.3 Experiments

Experiments were devised to evaluate what effect the proposed method has when different amounts of labelled data are available. We chose to investigate the task of multi-class deep grey matter segmentation on T_1-weighted MR images with 7 anatomical structures. This application was chosen for three main reasons. Firstly, it is a typical medical image segmentation task,

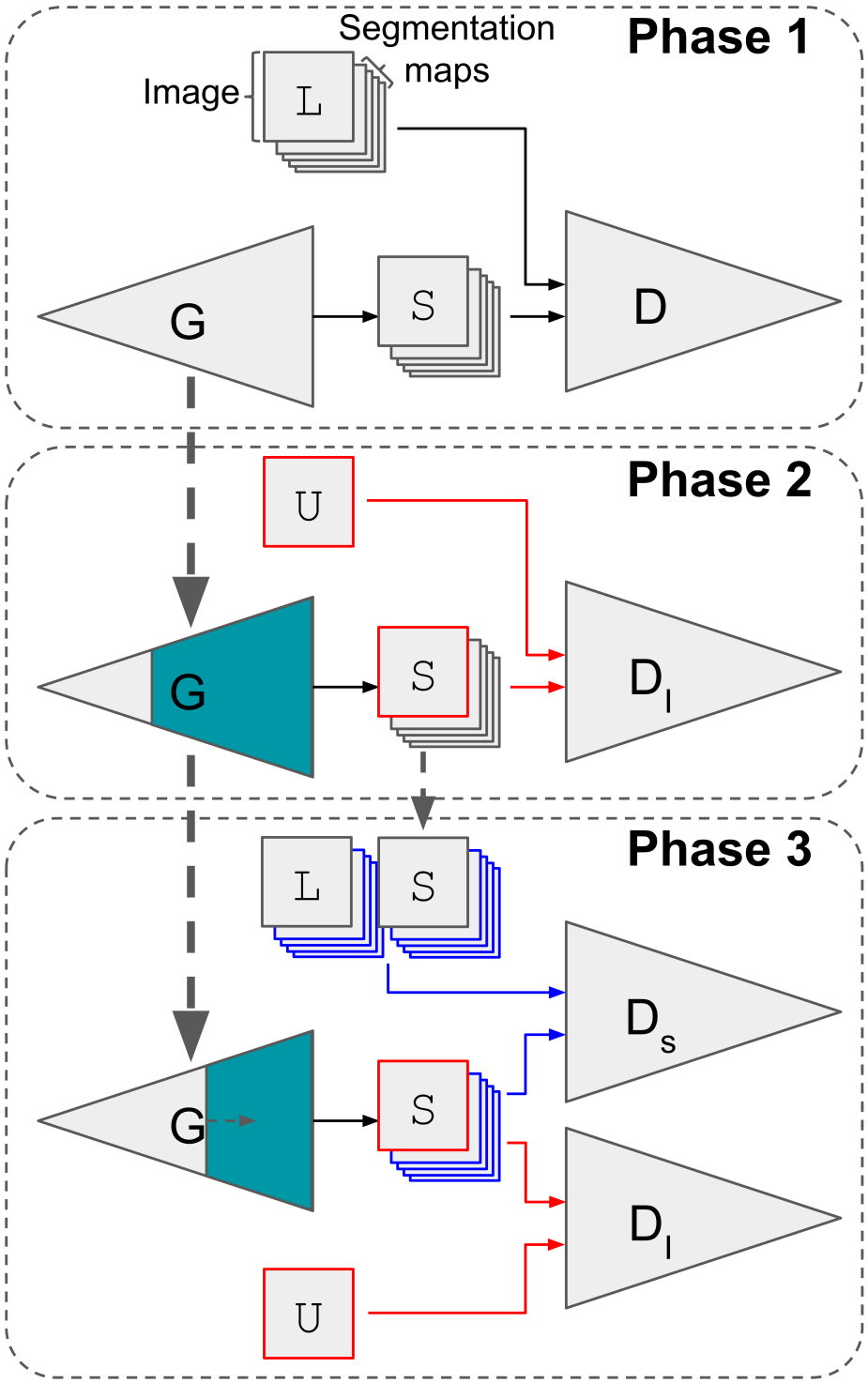


Figure 6.5: The three phases of GANsfer learning. *Phase 1* trains the whole network with generator (G) and discriminator (D) on labelled data (L) to produce synthetic data (S). *Phase 2* trains the early layers of the generator using unlabelled data (U) and a new image based discriminator (D_I). *Phase 3* reintroduces the later layers using a combination of both labelled data, unlabelled data and previously synthesised images. It uses combined feedback from the image based discriminator and a new segmentation based discriminator(D_S).

and as such, results on this task should provide an indication of performance on similar tasks. Secondly, data for multi-class problems are particularly time-consuming to manually annotate, and therefore provide a realistic use case for GANsfer learning. Finally, the relatively small region of interest means that the GANs required can be trained to the full resolution of the images in a reasonable amount of time, allowing for an extensive investigation with 5 fold cross-validation.

6.3.1 Data

For the labelled dataset we use data provided for the MICCAI 2013 Grand Challenge on Multi-Atlas Labelling[1]. This contains 35 images from the OASIS-1 dataset, manually annotated by Neuromorphometrics, Inc[2]. Each image is accompanied by clinical information including age, gender and Clinical Dementia Rating (CDR) - a 5 point rating scale for Alzheimer's Disease (AD) consisting of: healthy (CDR 0), very mild AD (0.5), mild AD (1), moderate AD (2) and severe AD (3). Of the 35 images, 5 are repeated from the same subject and are discarded. The remaining 30 images (Age: 18-90, median 25; Gender: 20F/10M; 29 healthy, 1 very mild AD) are affinely co-registered to a standard space with a 1mm isotropic voxel grid and intensity normalised to a zero-mean unit-variance across all non-background voxels, after which an 80-by-80-by-60px region of interest, defined in common space and covering the deep grey matter structures, is extracted. These images are then divided into 5 folds for cross-validation, each fold containing 24 training and 6 testing images, and with each image contributing 60 2-dimensional (2D) 80-by-80px axial slices.

Each slice has corresponding label information indicating the segmentation of the: Accumbens, Amygdala, Caudate, Hippocampus, Pallidum, Putamen and Thalamus, in the form of 7 separate binary image channels. As demonstrated in Figure 6.4, the GAN will learn to produce these channels from the same set of features which produce the MR channel. We, therefore, preprocess the segmentation channels to make them more closely correlated with the intensities within the MR channel. The aim here is to transform the binary segmentation channels into

[1] data available from `www.synapse.org/#!Synapse:syn3193805/wiki/217780`

[2] `www.neuromorphometrics.com`

continuous channels, with a value of 0 outside of the segmented region, and a value which is correlated with the corresponding MR intensities within the segmented region. We do this by replacing the regions within the segmentation mask with "residual" values, reflecting the contrast of the particular structure as compared to the surrounding White Matter (WM). To preprocess each segmentation channel, we transfer the pixel intensities from the MR channel into the corresponding region in the segmentation channel, and subtract an estimated WM intensity. These values are then inverted if necessary to remain positive and used as the segmentation channels for GAN training. This process is visualised in Figure 6.6.

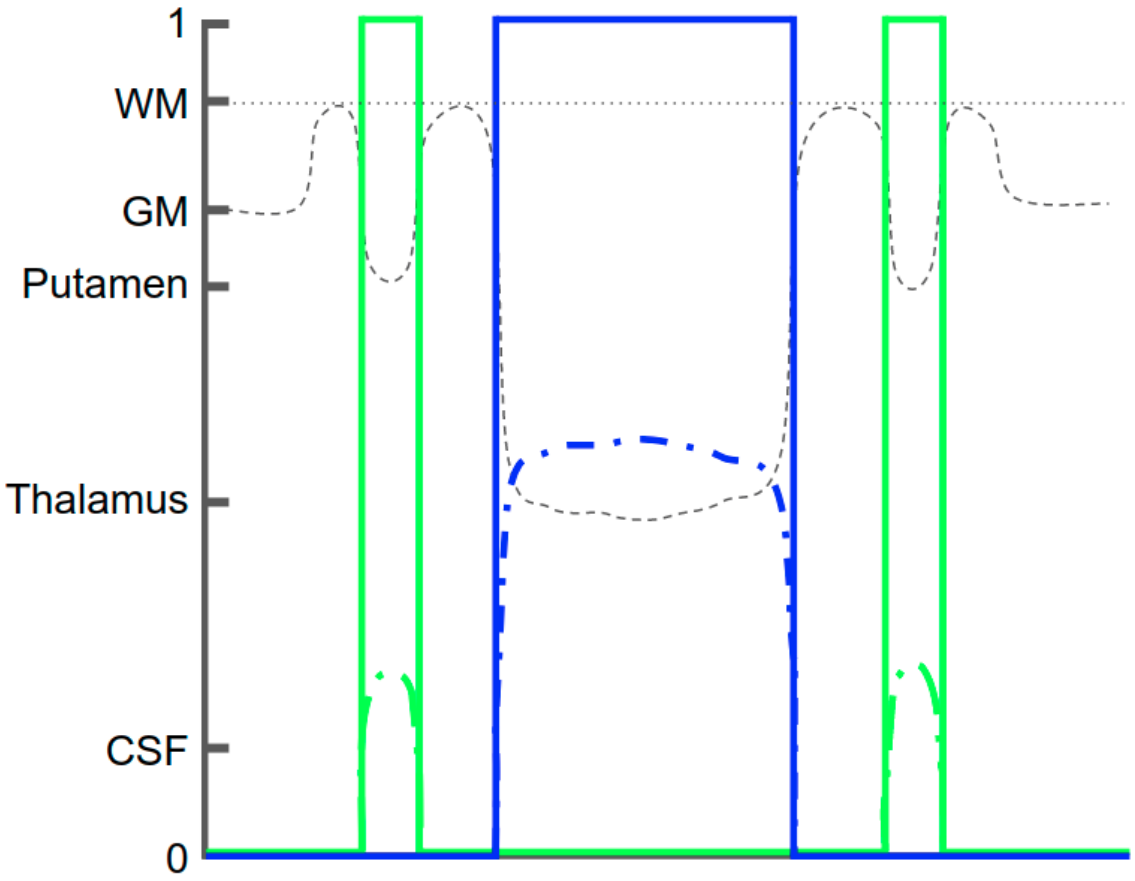


Figure 6.6: Visualisation of segmentation prepossessing steps. A hypothetical intensity profile showing the variation in MR image intensity along a line from left-to-right across an axial slice. The dotted line shows the relative image intensity, while the two solid coloured lines show the binary segmentation channels for the Putamen (green) and Thalamus (blue). The dot-dashed lines show the new values within the associated segmentation channels, calculated as the absolute difference between the MR intensities within these regions and the average WM intensity. The relative intensities of different structures are also shown (not to scale). Note that the new segmentation channels vary more smoothly, are correlated with MR channel intensity and are measures of the local contrast of each structure.

A key purpose of this is to remove the sharp edges present in the binary segmentation channels.

To generate these would require some of the feature maps in Figure 6.3 to have sharp edges, which could not also be used to generate the MR channel. This would lead to a potential decoupling of the parts of the network responsible for generating the MR and segmentation channels. To preserve the ability of the network to produce realistic segmentations after further training with unlabelled data, it is important that these processes are tightly linked. Towards this end, it is also important that the contrast levels in the segmentation channels are similar to the corresponding regions in the MR channel. All structures share a border with the WM, so it is, therefore, appropriate to use WM intensity as a base from which to calculate the contrast of each structure.

The unlabelled dataset consists of the entire OASIS-1 dataset. This contains 436 images (Age: 18-96, median 54; Gender: 268F/168M images; 336 healthy, 70 very mild AD, 28 mild AD, 2 moderate AD), of which 20 are repeated scans of the same healthy subject. These images are preprocessed in the same way as described above. This dataset has a much older age profile than the labelled dataset and contains many more examples of AD pathology. We hypothesise that by incorporating this older more pathological data into the GAN training process, the resulting network will produce more examples with features associated with old age and AD. This will then lead to more accurate segmentations of subjects with these features.

6.3.2 Postprocessing

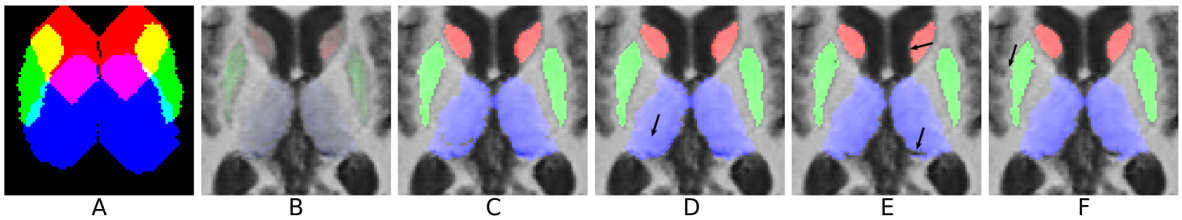


Figure 6.7: Post processing visualised on 3 segmentation channels. A) Mask derived from real segmentations. B) Masked image. C) Binarised segmentation channels. D) Holes filled. E) Intensity based threshold applied. F) Holes filled and spurs removed. Arrows indicate the effect of each step.

Since the GAN is trained to produce non-binary segmentation channels, a set of postprocessing

steps are required to binarise them and correct minor errors. These include small regions away from the expected location of a structure (usually within the Grey Matter (GM)) having a non-zero value in the corresponding segmentation channel, and synthesised segmentation channels extending beyond the boundaries of the structure they represent. First, each image is assigned a slice number through nearest neighbour comparison of the MR channel to the real training images. A 3-dimensional (3D) mask is then defined for each segmentation channel containing all points within 10mm of the union of all available labelled training images for that experiment. The appropriate slice from this 3D mask is then used to remove any obviously incorrect segmentations from each generated image. Each segmentation channel is then binarised and any resulting holes removed through morphological closing and hole filling. The MR intensity distribution within each structure is then calculated, with any pixels falling outside of the mean $+/-$ 2 standard deviations removed from the segmentation. Finally, any new holes are filled and morphological opening removes any small disconnected components and spurs. This whole process can be seen in Figure 6.7.

As well as post-possessing, the generated images are also filtered based on image quality. Whilst generated images were generally found to be of reasonable quality, the generator can occasionally produce unrealistic images. To filter these out, a score for each image defined as the minimum Euclidean distance between the MR channel of each generated image and any image from the full dataset is found. These are ranked and the generated images with scores above the 75^{th} percentile are removed. Unrealistic images were observed at a significantly lower rate than this (less than 5%), however, we use this very conservative threshold to guarantee no unrealistic images are kept, as removing realistic images is not detrimental since more can be generated at very little cost. Examples of the highest scoring images are shown in Figure 6.8.

6.3.3 Segmentation network

To assess the quality of the synthetic data, we propose to train a segmentation network with and without synthetic labelled data added to the available real labelled data. We use DeepMedic [Kamnitsas et al., 2017b] for this purpose. DeepMedic is a general purpose segmen-

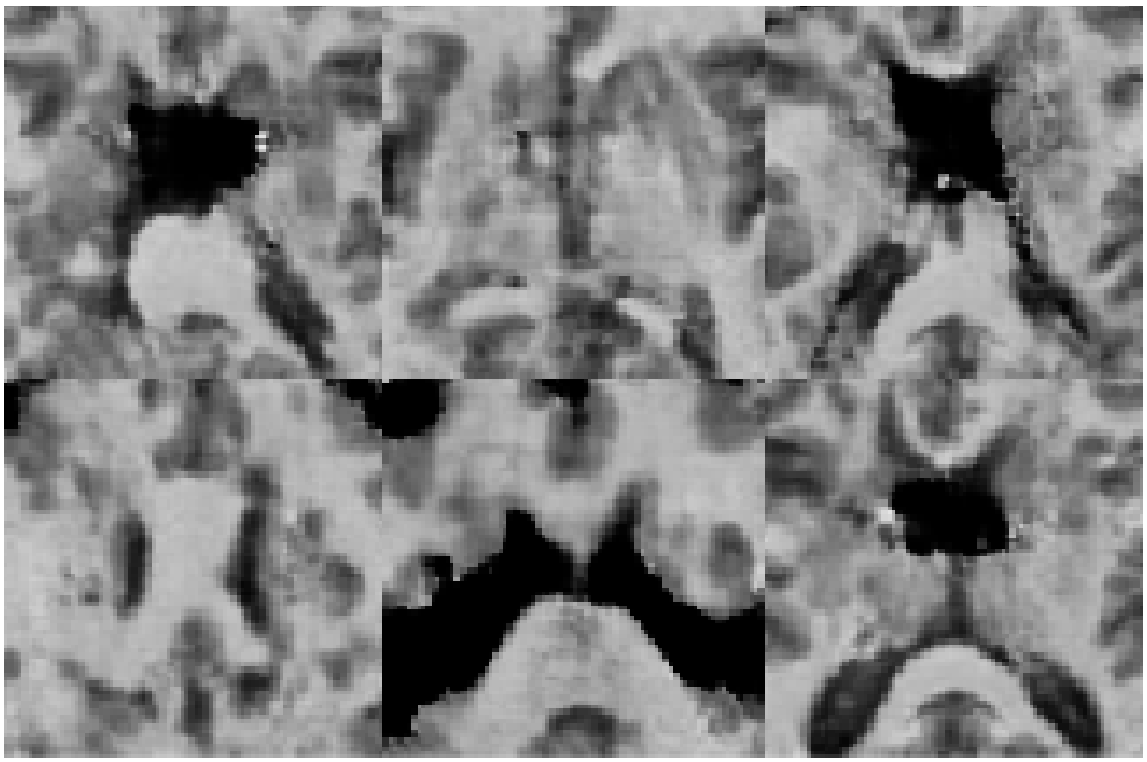


Figure 6.8: Examples of 6 unrealistic generated images removed from the dataset.

tation network which has been shown to give good results in a variety of segmentation tasks. We modify the default network architecture from 3D to 2D as described in [Kamnitsas et al., 2017b], and otherwise use default hyper-parameters and settings, including left/right reflection augmentation. Segmentations are evaluated using DSC (see 3.2.1). The primary metric used for comparison is an overall DSC, treating all 7 tissue classes as foreground, though we also examine the results for individual structures.

6.3.4 Impact of the quantity of available labelled data

For every learning task, there is a threshold beyond which additional labelled data provides negligible improvement. The goal of augmentation is therefore to lower the amount of data required to reach this optimal performance. Baseline experiments using 1, 3, 6, 12 and 24 labelled training images show no significant (two-tailed t-test, 5% significance level) difference between experiments using 12 and 24 images (see Figure 6.13 later for full results), indicating the that this optimal level of performance has been reached. The aim is, therefore, to achieve

results closer to this level using fewer labelled images and augmenting with synthetic data. Experiments were therefore performed by making 1, 3 and 6 labelled images available, taken from the pool of 24 allocated training images at each fold. Despite not expecting any significant performance increase on the baseline results when 12 and 24 labelled images are used, a subset of experiments is also performed at these levels to better understand the effects of synthetic data.

6.3.5 Impact of the quantity of synthetic data

The quality of synthetic data produced by the GAN will never be as high as real manually labelled images. We must, therefore, consider the relative exposure the segmentation network should get to the real and synthetic pools of training data. This is done by allowing the segmentation network to sample from each pool of data with a different probability, effectively allowing for different ratios between real and synthetic data to be explored. We consider 4 different ratios: 100:1, 10:1, 2:1 and 1:1. For example, a ratio of 100:1 means that for every 100 patches sampled from real data, 1 patch is sampled from synthetic data. Note that the pool of synthetic images available is effectively infinite, hence it is highly unlikely to sample the same image twice, while the pool of real images is relatively small, meaning the same region is likely to be sampled multiple times during training.

6.3.6 GAN training with large amounts of data

Early experiments indicated that the GAN training process produced higher quality images when trained on fewer (1, 3 or 6) labelled images than when more (12 or 24) images were available. Though more variation was observed when more images were used, the appearance would suffer (see Figure 6.9). This can be attributed to the GAN attempting to perform interpolation between exemplar images - behaviour which is not exhibited when few images are used, with the GAN generating images from around these modes with small differences. Once a sufficient amount of training images is provided, the GAN begins to attempt to produce images with

greater variation. This is normal and usually desired GAN behaviour, however, it is counter-productive in *phase 1*, where we desire high-quality images and variation is unimportant. This perceived loss of quality is confirmed when we perform segmentation experiments using only synthetic images after *phase 1*. Experiments on a single fold show an average overall DSC of 0.68, 0.73, 0.73, 0.68 and 0.64 when training on 1, 3, 6, 12 and 24 images respectively. To avoid this behaviour having a negative impact on the final results in the cases where 12 or 24 labelled images are available, the training data in these cases is further split into groups of 6 images, and 2 or 4 GANs respectively are trained using each group. After training, synthetic data from each GAN is combined and used to augment the real training data. Figure 6.10 shows an overview of the experimental setup across the 5 folds.

6.3.7 Full dataset analysis

A consequence of having limited labelled training data is that there is a similarly limited amount of test data upon which the trained segmentation models can be tested directly. While the 30 labelled images are sufficient to perform simple DSC based comparisons between different quantities of available data, they are insufficient for a deeper analysis. By only having one mild AD subject and 6 subjects with an age greater than 50, these 30 images do not provide a robust means of examining the effects on elderly and more pathological cases. We, therefore, perform further indirect evaluation on the unlabelled data by applying the trained segmentation models across the full dataset and using the volumes of the segmented structures as features to build a classifier to differentiate between cases of very mild AD and mild or moderate AD (CDR 0.5 and CDR 1 or 2).

After removing all repeated scans and training images, 287 healthy, 69 very mild AD, 28 mild AD and 2 moderate AD subjects remain available for analysis. Each image is segmented using five of the trained models (one per level of available data) and the volumes of each structure are extracted. These volumes form a 7-dimensional feature vector (one component for each of the seven tissue classes) for each subject, which are used to train a simple logistic regression classifier. Each experiment uses 5-fold cross-validation and is repeated 100 times. The observed

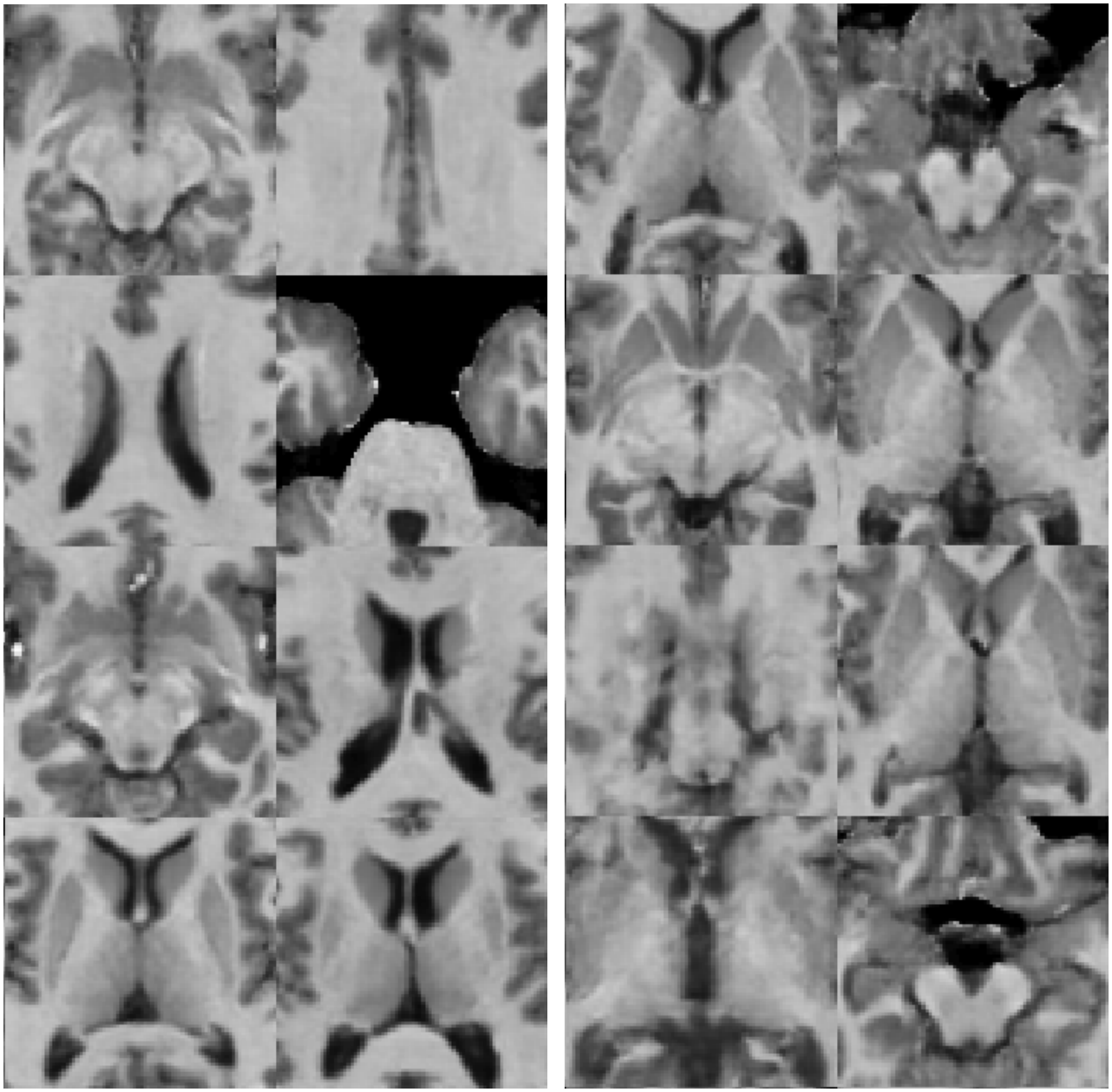


Figure 6.9: 8 random samples from the generator after *phase 1* training using 6 (left) and 24 (right) images. Despite having more training images, some images produced from 24 training images appear of low quality with a "dirty" appearance or unrealistic anatomy. Those produced from 6 training images are consistently of higher quality.

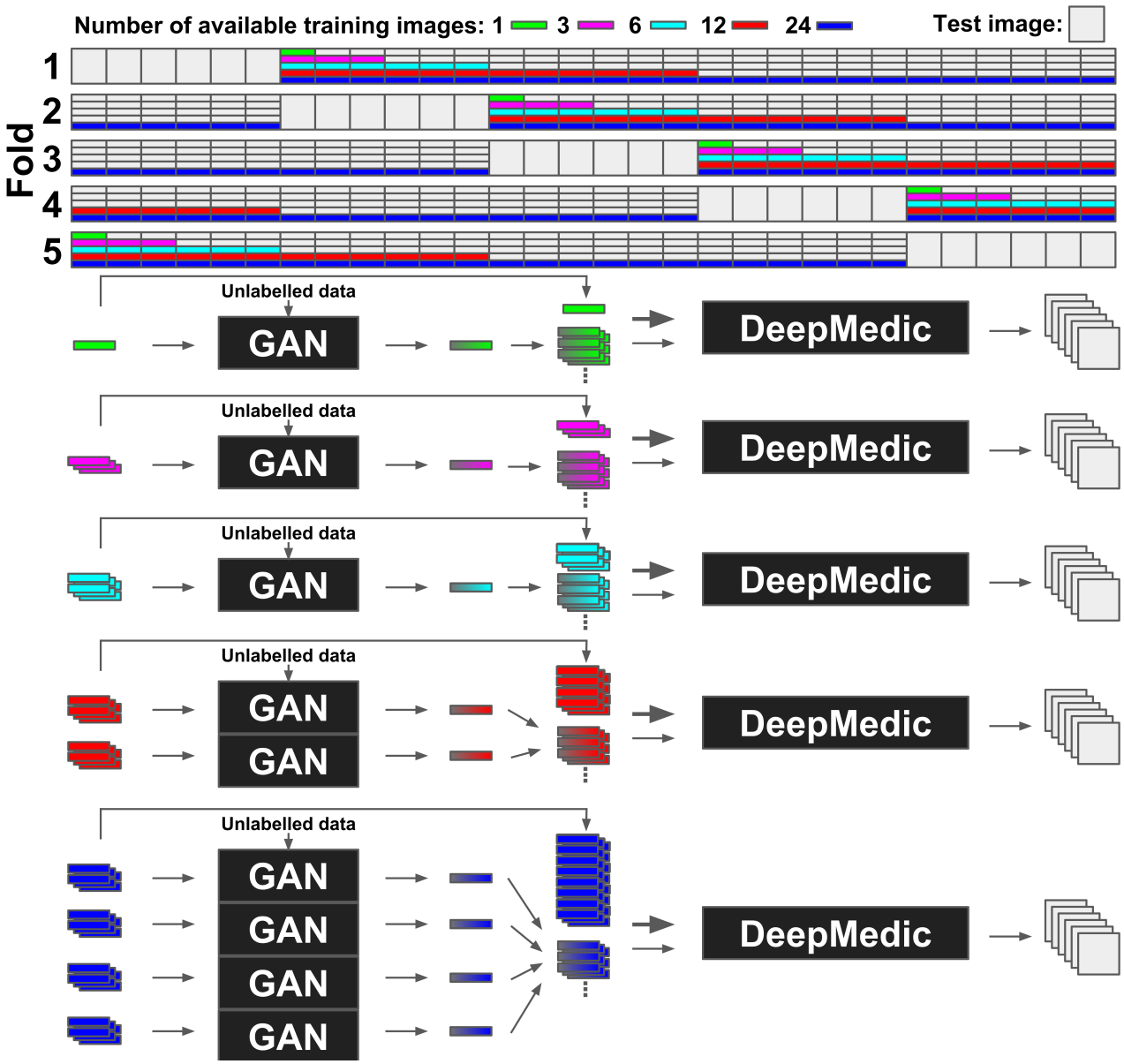


Figure 6.10: An overview of the experimental setup. At the top, the 30 labelled images are divided into training and test sets for 5 folds. For each fold, the training set is further divided to simulate cases where 1, 3, 6, 12 or all 24 images are available for training. Underneath, the process of training the required GANs and DeepMedic networks to investigate each level of available labelled data (1, 3, 6, 12 and 24, colour coded as above) for a single fold is shown. The available labelled and unlabelled data is first used to train a GAN using GANsfer learning. The generator is then used to create a synthetic dataset. A DeepMedic network is then trained by sampling (with varying probabilities) from the real and synthetic data, with the resulting model used to segment the 6 test images for that fold. Note that in the case of 12 (red) and 24 (blue) labelled images, multiple GANs are trained on blocks of 6 images, rather than training a single GAN on the all the images.

accuracy and Area Under the Curve (AUC) are calculated.

For these experiments, we use the models previously trained for one of the 5 folds, which was

chosen as its training set does not contain the AD subject and has the lowest average age of all folds (25). This choice of fold lets us explore the scenario where only labelled images for young and healthy subjects are available, yet we wish to segment images of older and pathological subjects. We also perform the same analysis using segmentations provided by Multi-Atlas-Label Propagation with Expectation-Maximisation based refinement (MALPEM) [Ledig et al., 2015] to allow for further comparisons. MALPEM is a 3D multi-atlas method to perform tissue segmentation and has been applied successfully in AD progression studies [Ledig et al., 2018b].

MALPEM offers significantly (two-tailed t-test, 5% significance level) greater DSC results than observed when using DeepMedic. This difference between the two methods is likely a result of MALPEM being a 3D method, and therefore having access to more contextual information. Baseline results using the two methods on a single fold show DeepMedic achieving an overall DSC of 0.80, 0.85, 0.87 when using 1, 6 and 24 labelled images respectively, with MALPEM achieving results of 0.80, 0.92 and 0.92 using the same subject atlases. The segmentations provided by MALPEM using all training atlases can, therefore, be used as a surrogate for manual segmentations. The agreement between these and the computed segmentations can be analysed under the assumption that a greater overlap with MALPEM segmentations would indicate greater accuracy. While not providing an absolute measure of performance, it does allow for further insights to be obtained making use of the full unlabelled dataset. Using the same structural segmentations as used in the classification experiment above, we compute the overall and per-class DSC using MALPEM segmentations as reference. We then examine how subject age and CDR classification affect the improvement seen when using synthetic data augmentation.

6.4 Results and discussion

6.4.1 Ablation and optimisation

To examine the impact of each stage of the proposed method we performed an ablation study where we measure results at the end of each phase of GANsfer learning. We also examine the impact of the segmentation channel pre- and post-processing, and the filtering of unrealistic synthetic images. Using a single fold and one available labelled image, we performed segmentation using the generator output at the end of each stage. We evaluated the synthetic data on its own and combined in a ratio of 2:1 real to synthetic data. The results of this study can be seen in Table 6.1.

Phase 1 is similar to the methods used in Chapter 5, and does not involve any additional unlabelled data. It is, therefore, reassuring to observe that we do see an improvement in segmentation performance. There is little difference when real data is used in addition to the synthetic data. This suggests that the generated images encompass all the relevant information from the real images, and are of sufficient quality to train from.

Table 6.1: **Ablation study:** DSC observed on a single fold using one labelled training image at different stages during the GAN training pipeline. Results are given using synthetic images produced by the GAN at the end of each training phase, with (+) and without real data. Results when using binary segmentation channels (i.e. no pre- or post-processing of the segmentation channels) are also shown with (BinCh/Filt) and without (BinCh/NoFilt) the filtering of unrealistic synthetic images. The overall DSC, DSC for each structure, and mean DSC across all structures are provided. Baseline results using no synthetic data are also shown for reference.

	Overall	Accum.	Amyg.	Caud.	Hippo.	Palli.	Putam.	Thal.	Mean
Baseline	80.1	46.1	56.5	80.0	57.7	78.0	81.4	84.0	69.1
Phase 1	81.6	57.2	56.2	81.7	66.9	77.5	81.7	86.6	72.5
Phase 1+	81.4	51.1	58.8	79.6	66.5	80.3	82.3	85.2	72.0
Phase 2	79.7	23.6	51.0	78.7	66.7	71.6	75.0	84.3	64.4
Phase 2+	82.0	48.5	59.7	81.2	67.8	81.0	79.5	86.8	72.1
Phase 3	79.5	44.2	55.0	80.4	70.5	69.2	74.8	85.1	68.4
Phase 3+	83.1	54.3	63.4	82.5	71.8	80.1	80.6	87.0	74.2
Phase 2&3	79.0	37.6	52.6	80.3	67.9	71.2	73.3	84.7	66.8
Phase 2&3+	83.9	56.1	61.9	83.6	74.3	80.5	81.4	87.6	75.1
BinCh/NoFilt+	80.7	40.3	58.9	76.9	60.5	76.7	80.5	86.6	68.6
BinCh/Filt+	82.9	47.8	61.5	81.6	63.0	76.8	83.6	87.4	71.7

After *phase 2*, using the synthetic data alone leads to worse results than observed after *phase 1*, but when combined with real data, it produces better results. This can be attributed to the synthetic data now containing additional information having been exposed to the unlabelled dataset. There is, however, also a reduction in image quality, leading to a worse performance when used on its own. This reduction in quality is addressed in *phase 3*, as evidenced by improved results compared to those after *phase 2*. The best results were then found by combining the synthetic data produced after *phase 2* and *phase 3*. This improvement is mostly driven by better hippocampal segmentation. Since the hippocampus is known to be affected by AD, it is possible that by using both sets of synthetic data, we include more examples of AD pathology. Alternatively, it could be that the segmentation network benefits from the additional variation after *phase 2*, even if the images are more unrealistic, and that some of this variation is lost during *phase 3*. Combining the two allows the network to be exposed to some additional variation, while also benefiting from the improved quality of the images produced after *phase 3*. All future experiments, therefore, use a combination of synthetic data from *phases 2 and 3*.

The effects of each phase of training can also be visualised. Figure 6.11 shows a tSNE visualisation [Maaten and Hinton, 2008] of training images, and synthetic images after each phase of

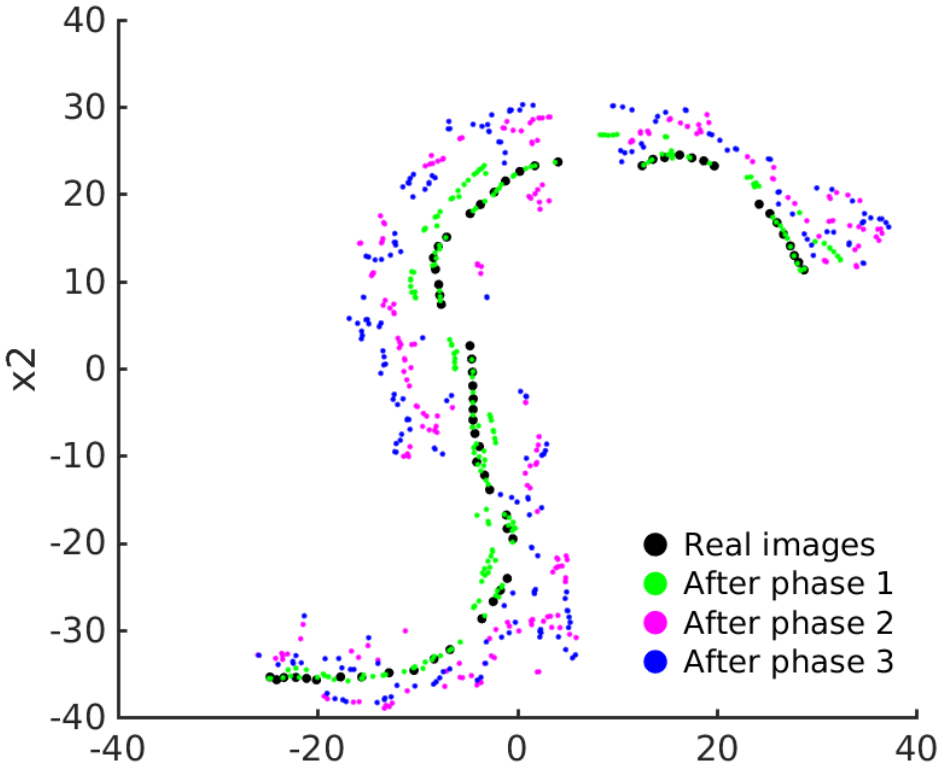


Figure 6.11: T-Distributed Stochastic Neighbour Embedding (tSNE) visualisation of training images, and the output from the same random selection of latent vectors after each training phase. The single training volume contributes 60 images and the output after *phase 1* follows these closely. Images produced after *phase 2* and *phase 3* are further away, indicating greater variability. Axes $x1$ and $x2$ correspond to the embedding coordinates found through tSNE.

training. A tSNE embedding allows us to visualise high dimensional data in a low dimensional space. Points which are close in the high dimensional space will also be close in the low dimensional embedding and vice-versa. This gives an indication of variation, but not quality. The further away the points corresponding to the synthetic images are from those corresponding to the real images, the more variation has been introduced. We are therefore looking to see points progressively move away from the real images after *phases 1 and 2*, and to have not moved closer again after *phase 3*.

Figure 6.12 shows sample output for two regions, at the end of each phase of training. These show that, whilst there is a reduction in image quality after *phase 2*, this is rectified in *phase 3* with no obvious loss of variability apparent in either Figure 6.11 or Figure 6.12.

Results when sampling real and synthetic data at different rates during DeepMedic training can

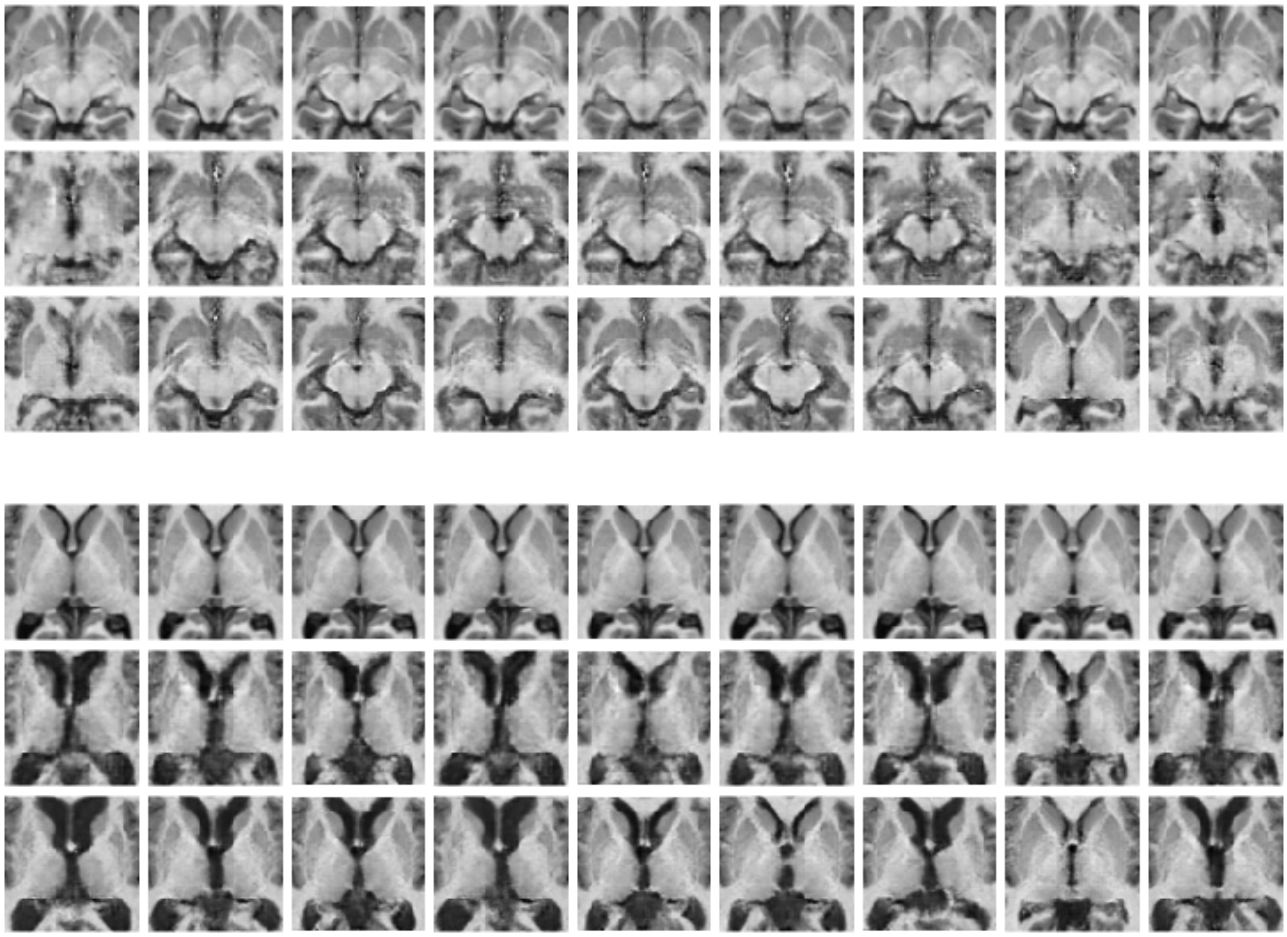


Figure 6.12: GAN output after each phase of training covering two example regions. For each region, 9 latent vectors were found which map to approximately the same image after *phase 1* (first row). The output from the same latent vectors were then generated after phases *2* and *3*. There is significantly more variation found after *phase 2*, at the cost of lower quality images (second row). An improvement in quality, while maintaining variability, can then be seen in the output after *phase 3* (third row).

be seen in Figure 6.13. The results clearly show that more synthetic data is beneficial when less real data is available. This is expected, as when more real images are available, more variation is already present in the training set, and therefore the additional synthetic data will have less impact, to the extent that when 12 or more real images are available, there is no evidence of improvement at any ratio.

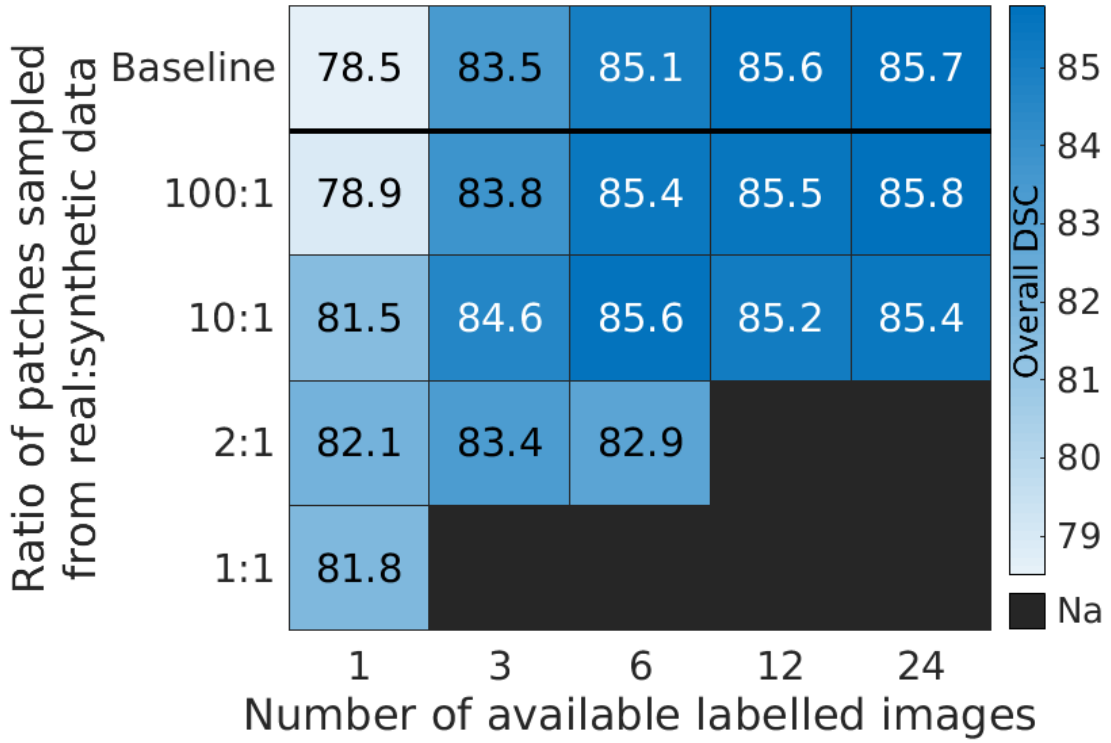


Figure 6.13: Observed DSC at baseline, and with different sampling rates of synthetic data during segmentation network training. There is a clear trend of more synthetic data being useful as the amount of real data is reduced.

6.4.2 Labelled dataset analysis

Using the optimal ratios found previously, we now examine the results across the labelled dataset. Figure 6.14 shows the distribution of observed DSC values with and without synthetic data at each level of available real data. Whilst we see significant improvements when 1 or 3 labelled images are available, neither of these is sufficiently large enough to score higher than using 3 or 6 images respectively. However, when 6 labelled images are available, we do see results which are not statistically different from those seen with 12 or 24 labelled images. It is reasonable to assume from the lack of improvement between baseline DSCs for 12 and 24 images that this is approaching the maximum DSC which can be achieved on this dataset using this segmentation algorithm.

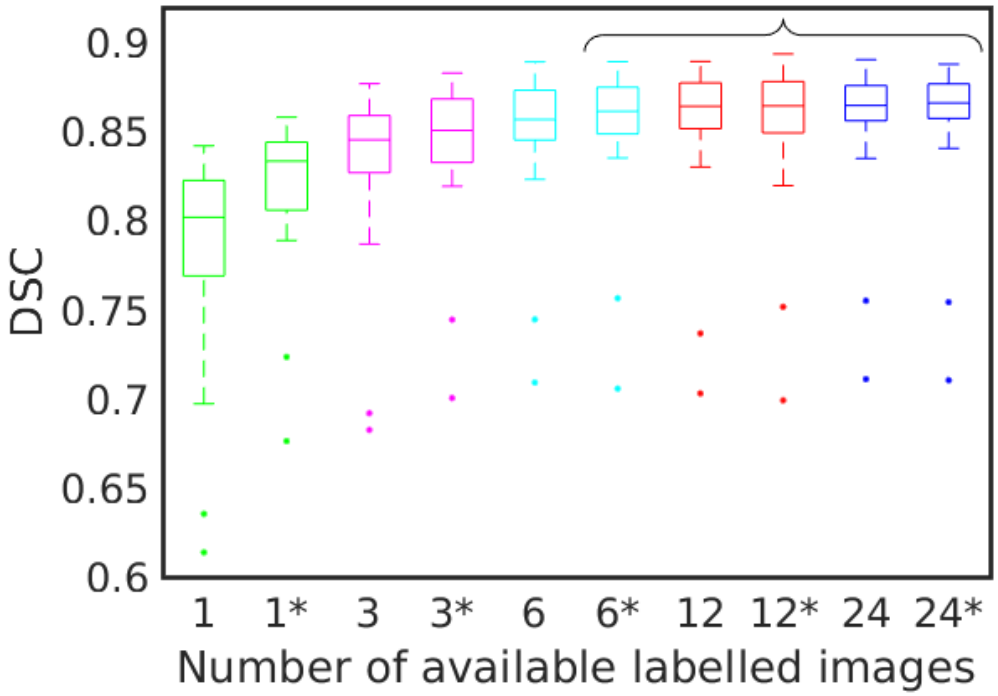


Figure 6.14: Box plots showing the distribution of DSC across all 30 images. Each coloured pair shows results with (*) and without synthetic data. Results within the bracket are not significantly different from each other at a 5% significance level calculated using a series of paired t-tests; all other results are significantly different from each other. The two outliers common to each experiment correspond to the eldest subject (lowest DSC) and single very mild AD subject (second-lowest DSC).

Observed DSC on each of the seven deep grey matter structures with and without synthetic data can also be seen in Figure 6.15. It is interesting to note that the largest improvements

in DSC can be seen in the hippocampus and amygdala, two structures which are known to be affected by AD. This suggests that exposing the system to more examples of anatomical variation in AD leads to an improvement in segmentation of these structures, even among a predominantly healthy cohort.

While the segmentation of most structures is improved, this is not the case for the Putamen. The reason for this is unclear and would require further investigation. However, it does have particularly high baseline scores with only a small improvement seen when using more real data, suggesting it is a relatively easy structure to segment with little room for improvement. We observed in Chapter 5 that in such cases where sufficient data is already available, adding synthetic data harms rather than helps. The performance on the Putamen may, therefore, be another example where the loss of image quality in the synthetic images is not sufficiently offset by the increased variation.

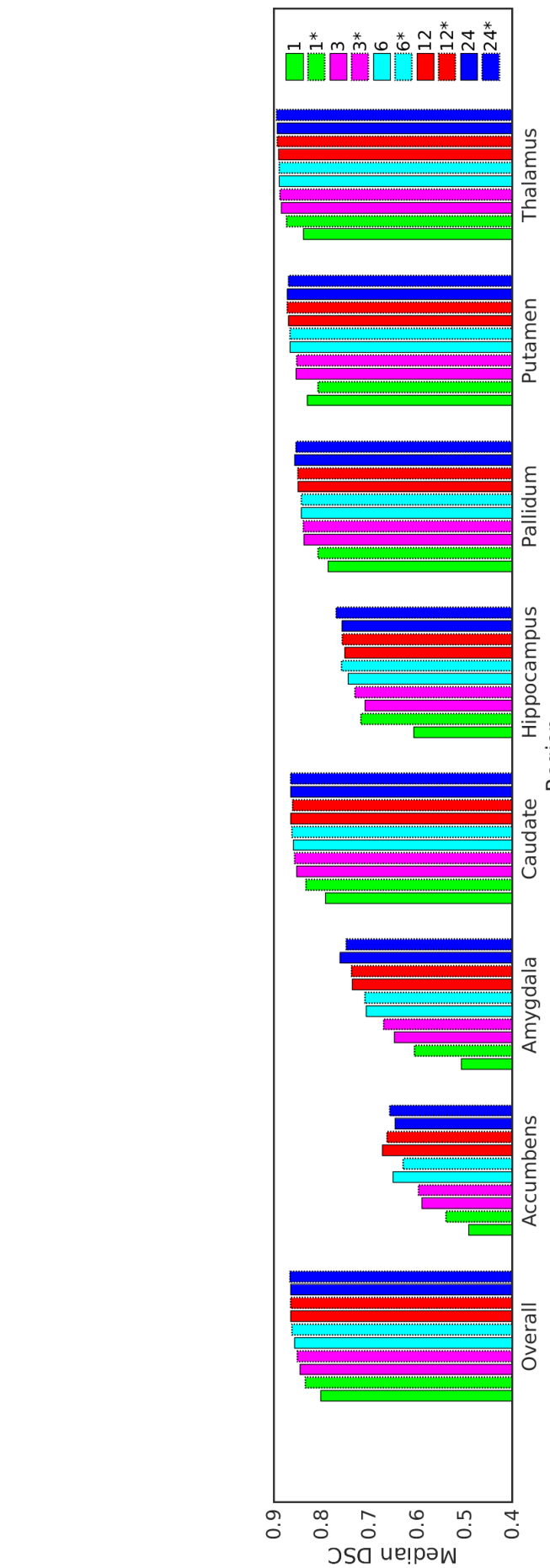


Figure 6.15: Impact of using additional synthetic data on segmentation accuracy for each of the seven structures. Each pair of coloured bars shows the difference between the baseline results (solid border) and results with the optimal amount of additional synthetic data (broken border).

6.4.3 Full dataset analysis

One benefit of using unlabelled data is that it extends the domain of training images from generally young and healthy to older subjects with more cases of AD. The impact of this can therefore only be fully measured on the full dataset. This was done by using MALPEM segmentations as a surrogate for ground truth labels. Figure 6.16 shows how the observed DSC varies with subject age, with and without synthetic data, along with a visualisation of the distribution of ages within the labelled and unlabelled dataset. It is clear that in all cases expected performance decreases as age increases, however, this effect is reduced when synthetic data is used. We can also see that in the cases where 12 and 24 labelled images are available there is little benefit to using synthetic data at the younger ages, agreeing with the results from the labelled data. However, there does appear to be a larger improvement at the older end of the age spectrum, beyond the range of ages provided by the labelled dataset. Figure 6.17 shows how the expected improvement in DSC varies with age across each structure. In each case, there is a clear trend showing that the further away from the ages present in the labelled dataset, the more benefit is given by using the synthetic data. This is particularly noticeable in the caudate. The proximity of the caudate to the lateral ventricles means that its location can vary significantly as the ventricles become enlarged by age, even if its volume is generally preserved. It is, therefore, a structure which could suffer from a spatial bias being learned when only a few healthy examples are provided, and is, therefore, a particular beneficiary of the additional anatomical variation introduced by the unlabelled data.

Figures 6.18 and 6.19 show a similar analysis, using CDR in place of age. Once again, the labelled dataset contains no examples of AD pathology, while the full dataset contains significantly more. Again we observe a clear loss of expected segmentation accuracy as CDR increases, but this effect is reduced when synthetic data is used.

The final experiment investigates whether using synthetic data leads to segmentations which were more useful at stratifying cases of AD, in particular, distinguishing cases of very mild AD (CDR 0.5) and mild/moderate AD (CDR 1/2). We also compare these results with those found when using MALPEM segmentation volumes calculated using all 30 unique labelled

images. Table 6.2 shows a clear benefit to both accuracy and AUC when synthetic data is used, where we observe a statistically significant improvement in 8/10 cases. In fact, using synthetic data increases the accuracy to a level that is close to that achieved by MALPEM, and leads to significantly better results when measured by AUC. In the case where 6 real images are available, the results show that the proposed method matches or outperforms MALPEM using 80% fewer labelled images.

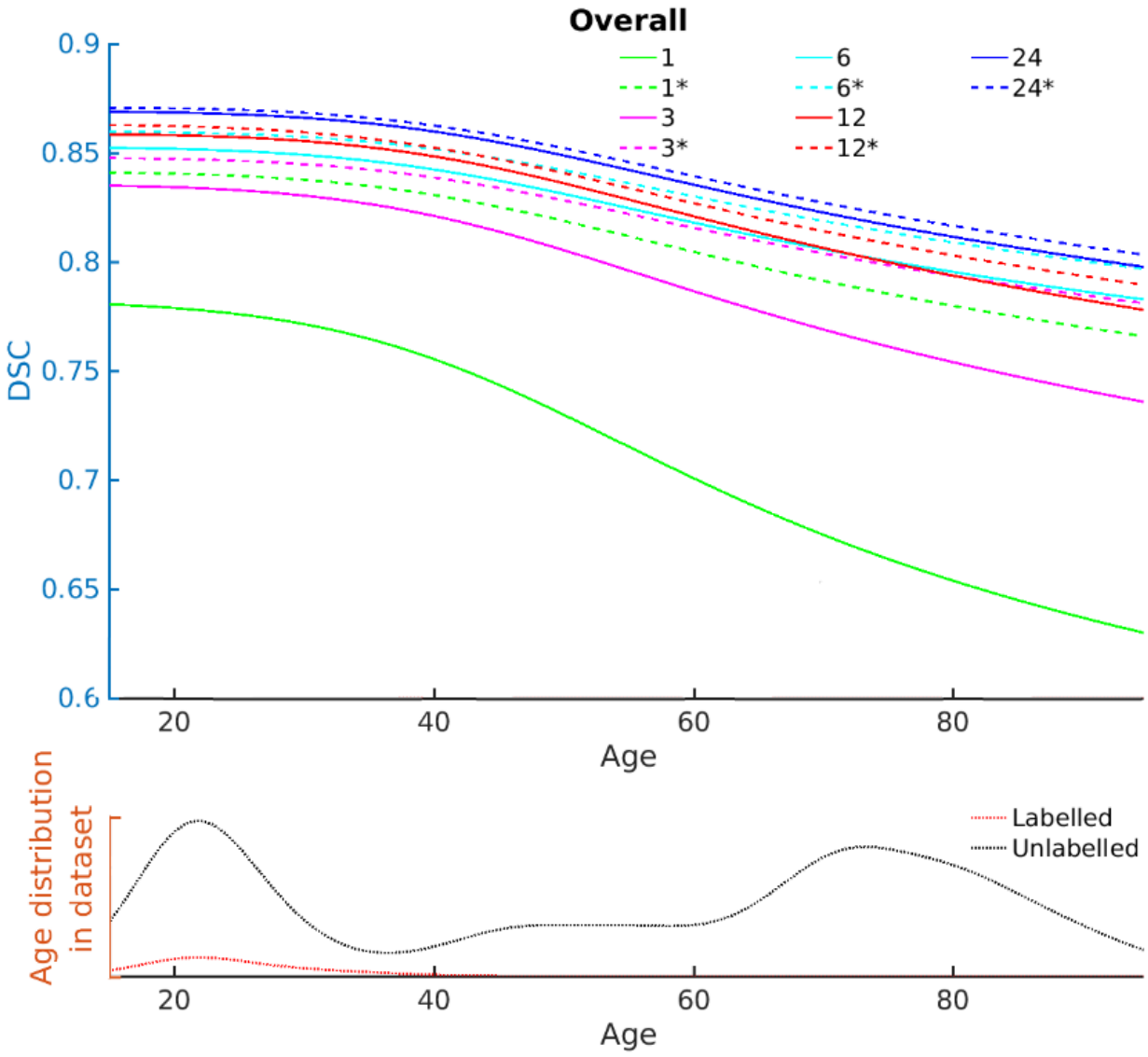


Figure 6.16: Overall DSC using MALPEM segmentations for reference at different ages. Results for segmentations computed with (*) and without synthetic data augmentation for each level of available labelled images (1,3,6,12,24) are shown. The relative age distributions for the full labelled and unlabelled datasets are also shown. All data is smoothed using kernel regression to highlight the overall trends.

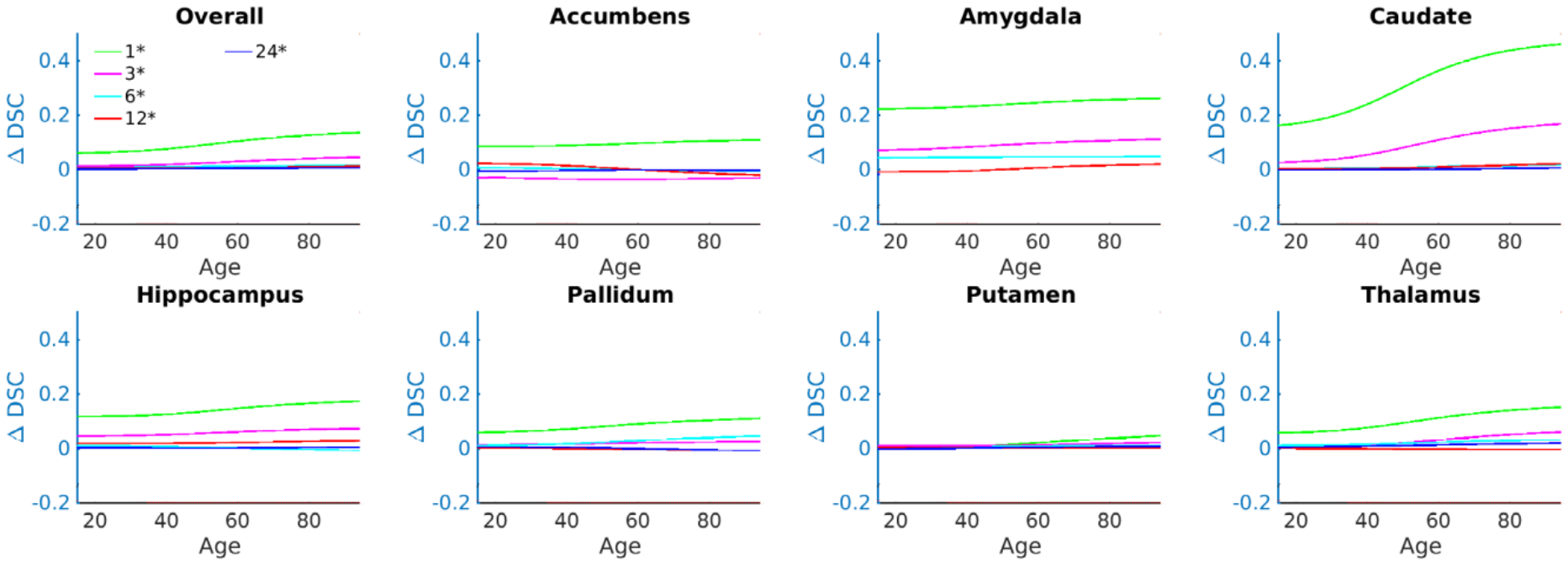


Figure 6.17: Difference in DSC using MALPEM segmentations for reference, overall and for each structure, at different ages. Pairwise differences in observed DSC between segmentations computed with and without synthetic data augmentation for each level of available labelled images (1,3,6,12,24) are shown. All data is smoothed using kernel regression to highlight the overall trends.

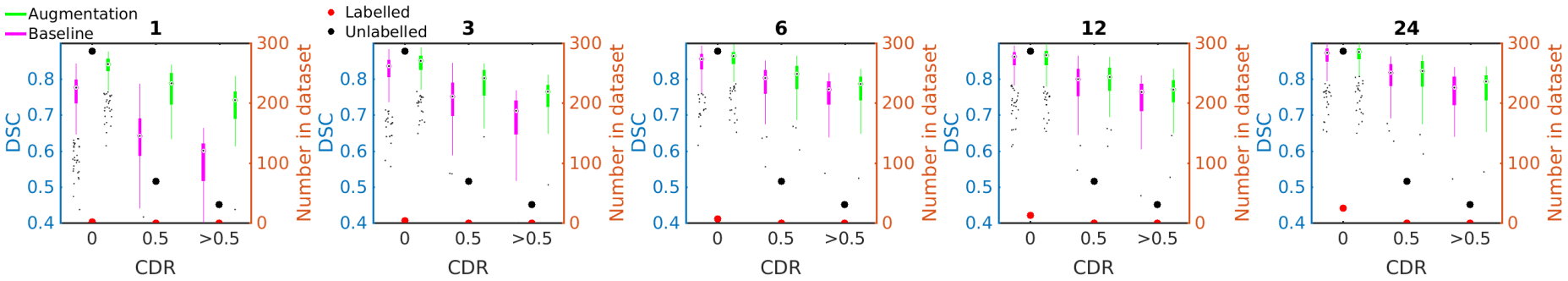


Figure 6.18: Distribution of overall DSC using MALPEM segmentations for reference for subjects with different CDR levels. The number of each group within the labelled and unlabelled datasets are also indicated.

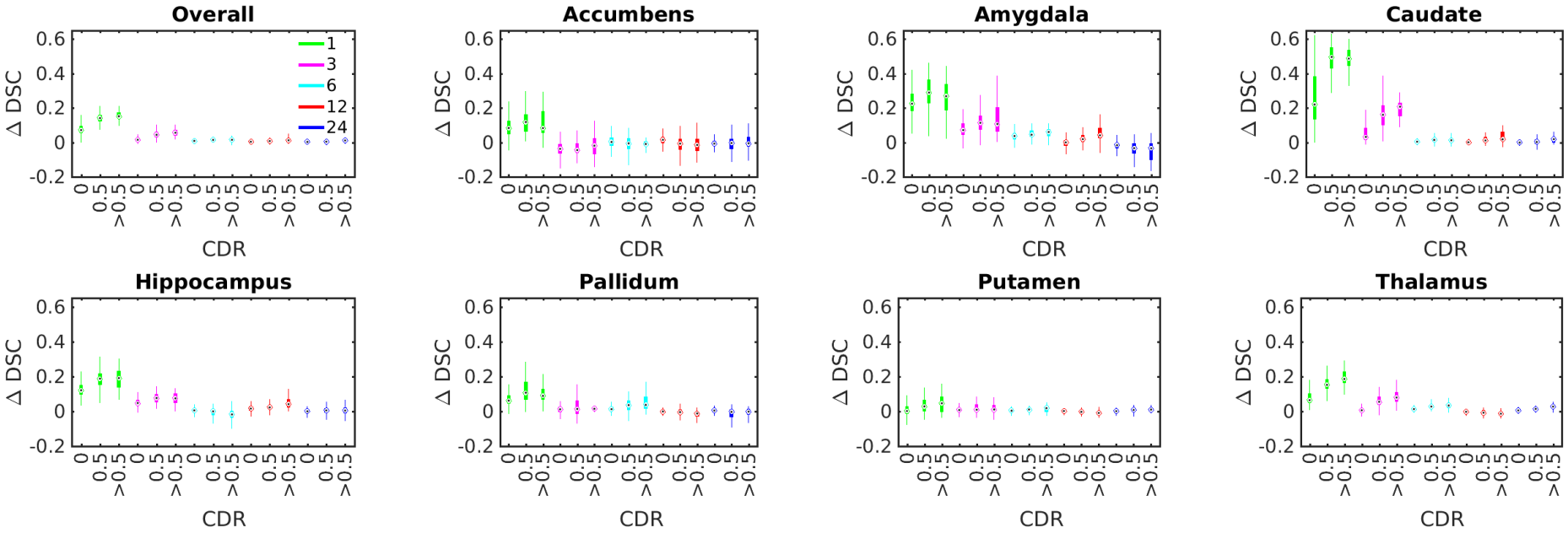


Figure 6.19: Distribution of paired DSC differences between results with and without synthetic data augmentation using MALPEM segmentations for reference. Overall results and results for each structure are shown for subjects with different CDR levels. Outliers are omitted for clarity.

Table 6.2: **CDR Prediction:** Accuracy and AUC metrics comparing the ability of segmentation volumes to differentiate between CDR 0.5 and CDR 1 or 2 subjects when computed with and without augmentation, and when using MALPEM (M). Results which are statistically different between corresponding baseline and augmentation results (2-tailed t-test, 5% significance level) are shown in bold. Results which are not significantly different from (†) and significantly higher than (*) the corresponding results using MALPEM (2-tailed t-test, 5% significance level) are also indicated.

	Accuracy						AUC					
Number of images	1	3	6	12	24	(M)	1	3	6	12	24	(M)
Baseline	64.3	62.9	66.0	63.0	63.3	67.6	56.4	47.9	55.0	58.7†	57.6†	58.5
Augmentation	**66.0**	**64.6**	**67.8**†	62.6	**66.6**		57.0	**56.7**	**65.6***	**60.0***	**66.7***	

6.4.4 Conclusion

In this chapter, we presented a method for incorporating unlabelled data into a segmentation network. We propose a novel GAN training procedure, similar to transfer learning, which allows a GAN to be trained on a mix of labelled and unlabelled data. The output of this GAN can then be used to augment the real training data. We performed a number of experiments, simulating cases where different levels of labelled data are available, and showed that the proposed method leads to the generation of labelled images with greater anatomical variation than was present in the labelled training data.

Three further sets of experimental results were presented. The first examined the effect of using synthetic data as measured by DSC on the labelled dataset. We observed that significant improvements can be made to DSC when small amounts of labelled training images are available, particularly in structures which are more strongly affected by AD. We also saw that by augmenting with synthetic data, 6 labelled images could be used to achieve the same results as 12 or 24 labelled images. Whilst this provides a useful indicator of performance, it does not evaluate one of the key aims of the proposed method - to extend the domain of training set from young and healthy to old and pathological. To evaluate this, we used a state-of-the-art

3D multi-atlas segmentation method to generate surrogate ground truth labels for the entire dataset. Using these, we showed how the expected DSC changes as age and AD diagnosis varies. We observed that segmentation results are substantially reduced once the test images fall outside of the domain of the training images either in terms of age or pathology. This effect was reduced by introducing synthetic data, leading to greater improvements in DSC for older subjects and those with AD. This confirms our hypothesis that exposing the network to information from unlabelled images of older and more pathological subjects through the synthetic data will lead to higher-quality segmentations of such subjects. Finally, the value of these higher-quality segmentations was demonstrated by showing that the stratification of AD between very mild and mild or moderate was significantly improved through the use of synthetic data.

Future work will involve applying the proposed method in different domains and further evaluating its benefits. There were some cases where using synthetic data led to worse results, particularly when ample training data was already available. This demonstrates that the synthetic images are no substitute for real images, implying either a lower image or segmentation quality. Investigating methods to improve image and segmentation quality will therefore also be a subject of future work. In addition, the ability to synthetically generate training data means that "bespoke" datasets can be created, for example, to address issues such as class imbalance. Alternatively, datasets could be created which contain additional examples of under-represented disease states or other demographics. Such an approach could even be taken during training, where an increased loss observed on a particular region of the test domain is reacted to by generating additional synthetic training data covering this region. Future work will therefore also involve investigating these applications.

During the previous two chapters we have demonstrated one application of GANs in medical imaging - for synthesising additional training data. However, there are many more potential uses for GANs which have yet to be fully explored. Of particular interest is the latent space learned by a GAN. So far in this thesis, we have simply sampled from this space to generate synthetic training images. However, these spaces have a number of interesting properties which we look to exploit in the remainder of this thesis.

Chapter 7

Modelling the Progression of Alzheimer's Disease in MRI Using Generative Adversarial Networks - Part A

7.1 Introduction

With an ageing population, neurodegenerative diseases are having an ever greater impact on society. The most common of which, Alzheimer's Disease (AD) has an estimated global cost of US$818 billion [Prince, 2015] with the 46 million sufferers in 2015 expected to rise to 131.5 million by 2050. This demonstrates a clear and urgent need for a global effort to tackle the disease, from earlier more accurate diagnosis, to patient care, to the search for new treatments.

Being able to accurately model the progression of AD is important for the diagnosis and prognosis of the disease, as well as to evaluate the effect of disease-modifying treatments. AD is associated with changes in the brain (described in more detail in Section 2.3.2). Primarily these changes involve increased atrophy, where parts of the brain shrink as cells die. This atro-

phy can be viewed on Magnetic Resonance Imaging (MRI) and is particularly apparent in the hippocampus, around the cortex and around the lateral ventricles [Seab et al., 1988]. Methods exist to quantify these changes such as the boundary shift integral [Freeborough and Fox, 1997] to measure brain volume change and automated hippocampus segmentation methods to measure hippocampal volume [Schuff et al., 2009]. Many models have been proposed demonstrating the development of these biomarkers, among others, including [Adak et al., 2004, Caroli and Frisoni, 2010, Doody et al., 2001]. Whilst there has been success in modelling the progression of AD related clinical biomarkers and image-derived features over the course of the disease, modelling the expected progression as observed by Magnetic Resonance (MR) images directly remains a challenge. Methods which model changes brain structures directly tend to focus on image-derived surface meshes [Thompson et al., 2001, Costafreda et al., 2011], not on the images themselves. In this chapter, we apply some recently developed ideas from the field of Generative Adversarial Networks (GANs) which provide a powerful way to model and manipulate MR images directly through a technique called image arithmetic. This allows for synthetic images based upon an individual subject's MR image to be produced expressing different levels of the features associated with AD. We demonstrate how the model can be used to both introduce and remove AD-like features from two regions in the brain, and show that these predicted changes correspond well to the observed changes over a longitudinal examination. We also propose a modification to the GAN training procedure to encourage the model to better represent the more extreme cases of AD present in the dataset.

The development of models such as this could have several applications. First, they would allow clinicians to measure the progression of the disease in a patient against a predicted path, and provide visual feedback to the patient and their family. They could also be used to measure the impact of drug trials where actual disease progression can be compared to a predicted intervention-free progression by comparing two images directly, as opposed to image derived measurements. Finally, they could lead to a powerful tool for patient motivation. For example, showing a patient a predicted progression with and without changes lifestyle, such as diet, could provide additional encouragement to change their behaviour.

This chapter is divided into two parts. Here in Part A, we investigate the ability for the latent

space learnt by GANs to create accurate predictions of AD progression. The work presented in Part B of the chapter re-implements and extends the methods presented in Part A using a more recent network architecture, unavailable at the time of the original work.

7.1.1 Generative Adversarial Networks

GANs, first proposed by [Goodfellow et al., 2014], have seen a lot of attention recently within the field of computer vision. Many developments on the original architecture have been proposed [Arjovsky et al., 2017, Tolstikhin et al., 2017, Zhang et al., 2017, Berthelot et al., 2017, Radford et al., 2015, Gulrajani et al., 2017, Zhao et al., 2016], though they all follow the same basic approach. Below is a brief summary of a GAN's function, whilst a more detailed review can be found in Section 3.1.1.

The goal of a GAN is to learn to generate synthetic samples from an unknown distribution for which a set of real samples are known. For image tasks, this means being able to generate synthetic images with the same characteristics as a training set of real images, ideally to the degree that the synthetic images are indistinguishable from the real images. This goal is accomplished through the playing of a game between two adversaries, the generator **G** and discriminator **D**. In this iterative game, the goal of **G** is to learn a mapping from a low dimensional random vector seed **z** to a high dimensional image such that **D** cannot distinguish these synthetic images from the real training images. **D**, therefore, has the goal of learning to distinguish the synthetic from the real images. After each round, **G** is updated according to the loss observed by **D** on the latest batch of synthetic images. As the game plays out, **G** becomes more sophisticated, eventually producing images with the desired characteristics of those in the training set.

7.1.2 Image Arithmetic

One of the key features of GANs which we exploit in this chapter is the idea of image arithmetic, where the characteristics of an image can be modified through simple arithmetic in a latent

space. With a fully trained generator, the random seed $\mathbf{z}$ can be considered a latent encoding of the resulting image $\mathbf{G}(\mathbf{z})$. The manipulation of an image's latent encoding has predictable and logically consistent effects on the resulting image. The classic example of this appears in [Radford et al., 2015] where the authors train a GAN on a dataset of faces. They proceed to show that by taking the latent encoding of an image of a man wearing glasses, subtracting the encoding of an image of a man without glasses, then adding the encoding of an image of a woman. The resulting encoding, when mapped back to image space, yields a woman wearing glasses. In addition, they also demonstrate the continuity of the latent space by making small modifications to the latent representations and showing that these corresponded to slightly different faces and glasses of differing tints in image space.

The goal of the work described here is therefore to generate subject-specific predictions of AD progression by isolating the latent encoding of the features which correspond to AD, and using this to introduce or remove these features from real images. This is much like how the authors of [Radford et al., 2015] isolate the features which correspond to glasses and add this to the features of another face. In addition, the continuity of the latent space allows us to introduce different amounts of these features by adding multiples of the isolated AD encoding. This allows for a series of images showing the progression of AD pathology to be generated.

7.2 Method

The Wasserstein Generative Adversarial Network (WGAN) [Arjovsky et al., 2017] framework was used as the basis for our experiments due to its simple and extensible formulation, and robustness. The same framework and hyper-parameters as described in [Arjovsky et al., 2017] were used. For the generator, we used a 256-dimensional input vector followed by 5 convolutional layers with 1024, 512, 256, 128 and 1 (output) feature maps respectively. Convolutions were performed with a stride of 2 to reach the desired output image size (64-by-64px). The first convolutional layer used a 4-by-4 kernel, with subsequent layers using 3-by-3 kernels. As usual, the discriminator architecture was set as the reflection of the generator with the final layer

mapping to a single output. We modify the training procedure to address some deficiencies we observed when using the standard framework.

7.2.1 Example re-weighting

A reduced ability to generate images with more extreme atrophy was observed during early experiments. While similar to the phenomenon of "mode collapse" sometimes seen when training GANs, where the generator maps any input to approximately the same output, what we observed was the GAN being unable to generate images from the extremes of the distribution, despite learning a smooth manifold elsewhere. We hypothesised that this was due to a lack of such examples in the training data and the tendency of a GAN to generate more "average" images - the GAN will not waste some of its limited source of entropy in explaining variation it rarely sees. To address this, we adopt a training data re-weighting schema similar to that seen in the AdaBoost [Freund and Schapire, 1996] algorithm. A similar idea was proposed in AdaGAN [Tolstikhin et al., 2017], where an ensemble of GANs are used. However, this would be inappropriate for our use as having an ensemble of GANs would not allow a single latent encoding for AD to be isolated.

Throughout training, during epoch t, each real image $\mathbf{I}$ in the training dataset $\mathbf{T}$ has an associated weight $w_{\mathbf{I},t}$ and discriminator loss $l_{\mathbf{I},t}$, with the loss for each batch $\mathbf{B}$ calculated as $\sum_{\mathbf{I}\in\mathbf{B}} w_{\mathbf{I},t} l_{\mathbf{I},t}$. After each epoch, each $w_{\mathbf{I},t}$ is updated according to the following rule:

$$\epsilon = \sum_{\mathbf{I}\in\mathbf{T}} w_{\mathbf{I},t} l_{\mathbf{I},t},\ \alpha = \frac{1}{2}\ln(\frac{1-\epsilon}{\epsilon}),\ w_{\mathbf{I},t+1} = w_{\mathbf{I},t} e^{-\alpha l_{\mathbf{I},t}}, \tag{7.1}$$

with $w_{\mathbf{I},0} = \frac{1}{|\mathbf{T}|}$ for all $\mathbf{I}$. Note that since Equation 7.1 comes from the discrete version of AdaBoost, in which the loss is expected to be binary (i.e. correct vs incorrect classification), it is necessary to binarise the discriminator output. This was performed by applying a threshold of 0, which was found to be a reasonable decision boundary in the presence of batch normalisation in the discriminator).

This has the effect of increasing the weighting of those real images which are misclassified (considered synthetic) by the discriminator. This forces the discriminator to better represent the more extreme parts of the image distribution, which will, in turn, encourage the generator to produce images from these regions. Figure 7.1 shows the relative weights of a selection of training images during the early stages of training. During this period, 6 of these images are identified as frequently misclassified and their weights increase until the discriminator learns to accept them as real, upon which their weights decrease again.

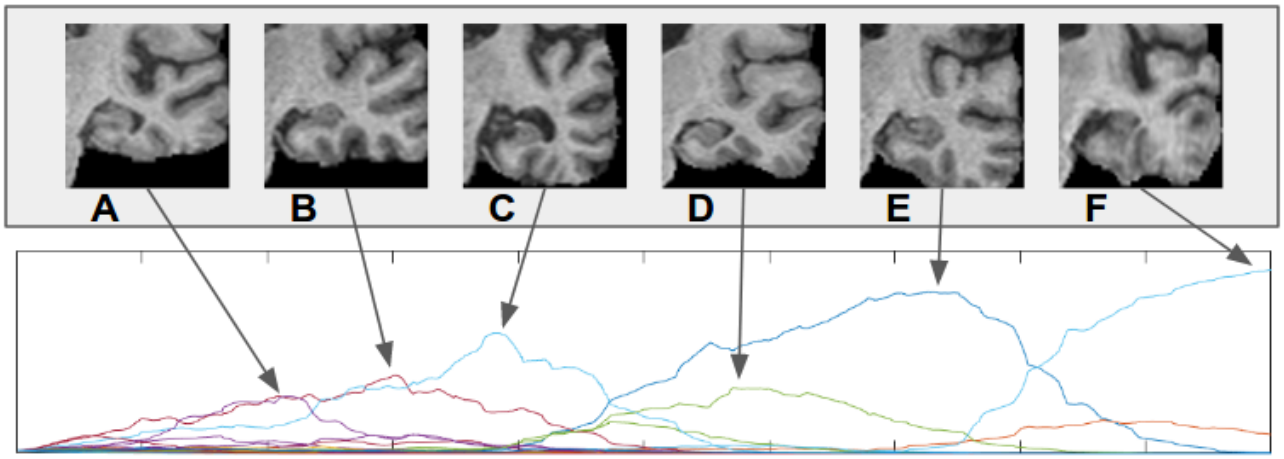


Figure 7.1: Bottom: Graph indicating the relative weight of 64 randomly selected images over the first 1000 iterations of training on a patch showing the hippocampus. Top: The 6 images with the highest weights. Each image has one or more unusual properties which place them towards the extremes of the image distribution. A,B&D: "Flattened" temporal lobe. C&F: Considerable atrophy. E: Possible artefact in the top right.

7.2.2 Encoding an image

Once a GAN has been trained, the generator $\mathbf{G}$ provides a mapping from latent encoding $\mathbf{z}$ to an image. However in order to manipulate the latent encoding of real images, a method to map from real image $\mathbf{I}$ to optimal latent encoding $\mathbf{z^*}$ is required, such that $|(\mathbf{G}(\mathbf{z^*}) - \mathbf{I})|^2$ is minimised. One solution is to train a third network to map from image to latent space, trained on batches of generated images and their corresponding encodings. Since the discriminator is already a trained feature extractor, it is possible to avoid training a new network by instead re-purposing the discriminator $\mathbf{D}$ by modifying the output layer to be of size $|\mathbf{z}|$, and to retrain $\mathbf{D}$ such that $|(\mathbf{D}(\mathbf{G}(\mathbf{z})) - \mathbf{z})|^2$ is minimised. $\mathbf{z^*}$ can then be found extremely quickly (<0.1

second) through a single forward pass through the retrained discriminator.

An alternative suggested in [Yeh et al., 2016] and used in [Schlegl et al., 2017] is to find $\mathbf{z^*}$ for a given $\mathbf{I}$ using a gradient descent approach. We use a similar approach whereby $|(\mathbf{G}(\mathbf{z}) - \mathbf{I})|^2$ is minimised by iteratively updating $\mathbf{z}$ so that it approaches $\mathbf{z^*}$. The main difference from the method in [Yeh et al., 2016] is the omission of a perceptual loss term, as such a loss would penalise the reconstruction of images from the extremes of the distribution, such as those with considerable atrophy.

This approach can be accomplished simply by adding a layer to the beginning of $\mathbf{G}$. This layer is designed such that it outputs the pointwise multiplication of its input with a set of weights $\mathbf{w}$. By setting the inputs to a vector of ones of size $|\mathbf{z}|$, the output of this layer, and therefore the input to $\mathbf{G}$, is $\mathbf{w}$. The optimal values of $\mathbf{w}$ can now be learned such that $|(\mathbf{G}(\mathbf{w}) - \mathbf{I})|^2$ is minimised, with $\mathbf{z^*} = \mathbf{w}$ at the end of training. By setting the learning rate to 0.05, convergence is reached in $\sim$20 seconds on a Tesla K80 GPU.

7.2.3 Isolating the visual appearance of AD using latent encoding

The latent encoding for AD can be found by subtracting the average latent encoding of a set of images corresponding healthy subjects from those corresponding to AD subjects. The resulting difference can then be added or subtracted from the latent encodings of real images to add or remove the features of AD.

7.2.4 Experimental Setup

GANs were trained on examples of images with different levels of AD. These images were taken from the ADNI dataset which contains images for over 1000 subjects with different levels of AD; from Normal Controls (CN), through Mild Cognitive Impairment (MCI), to AD. Most of these subjects have followup scans, with many showing a progression in the disease. This results in thousands of images containing subjects at different stages of the disease. When

training a GAN, it is important to have many examples of anatomical variation across the full spectrum of the disease. With this in mind, we disregard the disease classification and whether the image is a baseline or follow-up scan, so as to learn as much variation as possible. As a result, training was performed on 6220 images. Each image was brain-extracted using PINCR-RAM [Heckemann et al., 2015], bias-corrected using the N4 [Tustison et al., 2010] algorithm, affinely co-registered to a 1mm isotropic Montreal Neurological Institute (MNI) template using MIRTK[1] and intensity normalised so as to have a zero mean and unit standard deviation. A further 1000 images were preprocessed and kept separate for use in evaluation.

Two 64-by-64px regions were chosen for our experiments: A full resolution (1mm isotropic) region in a coronal slice showing the hippocampus and temporal lobe, and an in-plane down-sampled (3-by-3-by-1mm) transverse slice showing the lateral ventricles. To double the number of training images available for the first experiment, regions from both hemispheres were taken, with the images from the left side reflected and grouped with those from the right.

Four GANs were trained with an original WGAN and WGAN with re-weighting (WGAN+RW) trained at each region. In all experiments, the size of $\mathbf{z}$ was set to 256. All GANs were trained for 1100 epochs, well beyond the level of apparent convergence, taking approximately 10 hours on a Tesla K80 GPU.

A set of 1000 additional images were reconstructed with each method, with a sample of results shown in Figure 7.5. WGAN+RW was found to lead to visually more accurate reconstructions in cases of high atrophy or other abnormality, with little difference between the methods seen otherwise. The optimal latent encodings for each image using WGAN+RW were therefore calculated for 784 CN, 1154 MCI and 434 AD images from the training data. The latent encoding for AD, $\mathbf{z}_{AD}$, was then found as described previously. Images corresponding to $\mathbf{G}(\mathbf{z^*} + \gamma \mathbf{z}_{AD})$ were generated at both locations for a CN subject showing the addition of the features of AD where $\mathbf{z^*}$ indicates the optimal latent encoding of the original image found through gradient descent and γ controls the amount of these features introduced (Figure 7.7). Additionally, longitudinal images of an AD subject were predicted by a) Starting at the final image and removing

[1] https://github.com/BioMedIA/MIRTK

AD features, and b) Starting at the baseline image and adding AD features (Figure 7.8).

7.3 Results and Discussion

7.3.1 GAN synthesis

Samples of images generated by the GANs can be seen in Figures 7.2 and 7.3. These were produced using the WGAN+RW training schema. The generated images appear realistic with no obvious flaws and cover a range of atrophy levels and patterns.

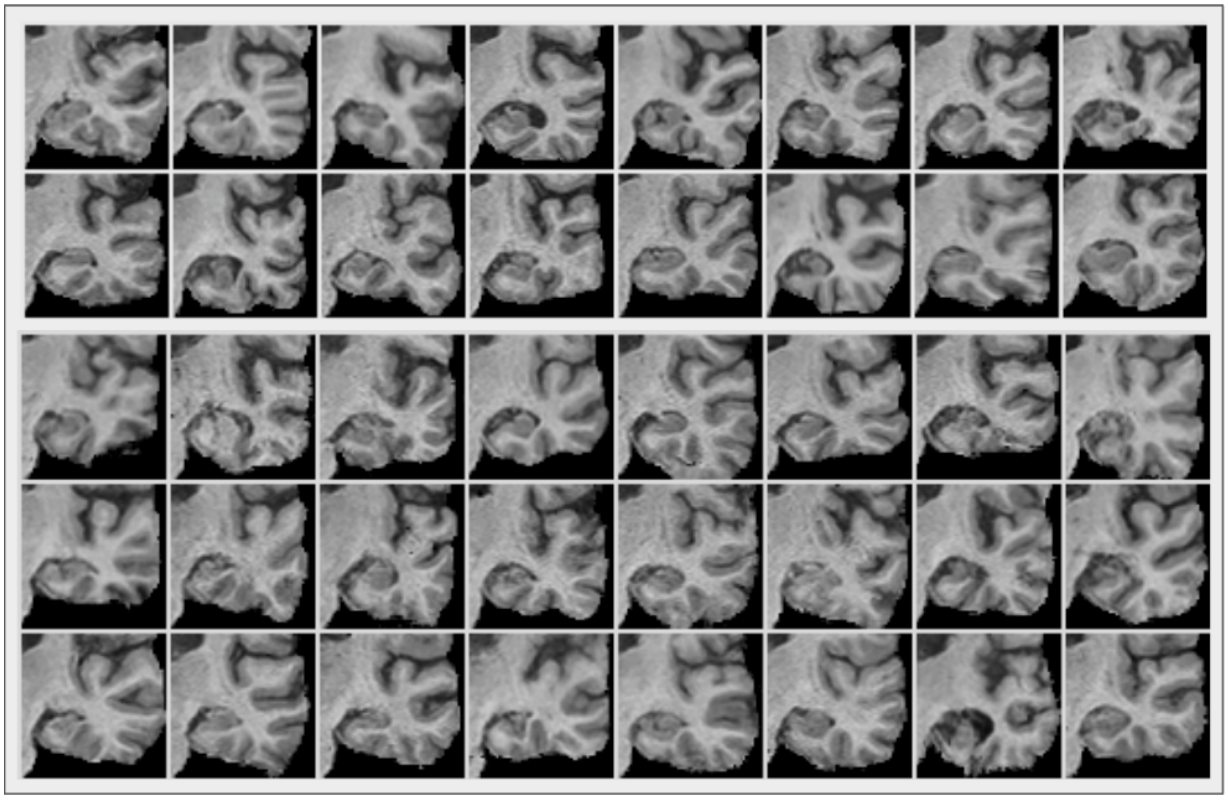


Figure 7.2: A random sample of synthetic images (bottom) of the hippocampus produced by the GAN, with a selection of real images (top) for comparison.

7.3.2 Encoding and reconstruction

Figure 7.4 shows reconstructions using both the retrained discriminator and iterative approaches to encode real images. The results after repeated encoding and reconstruction cycles were

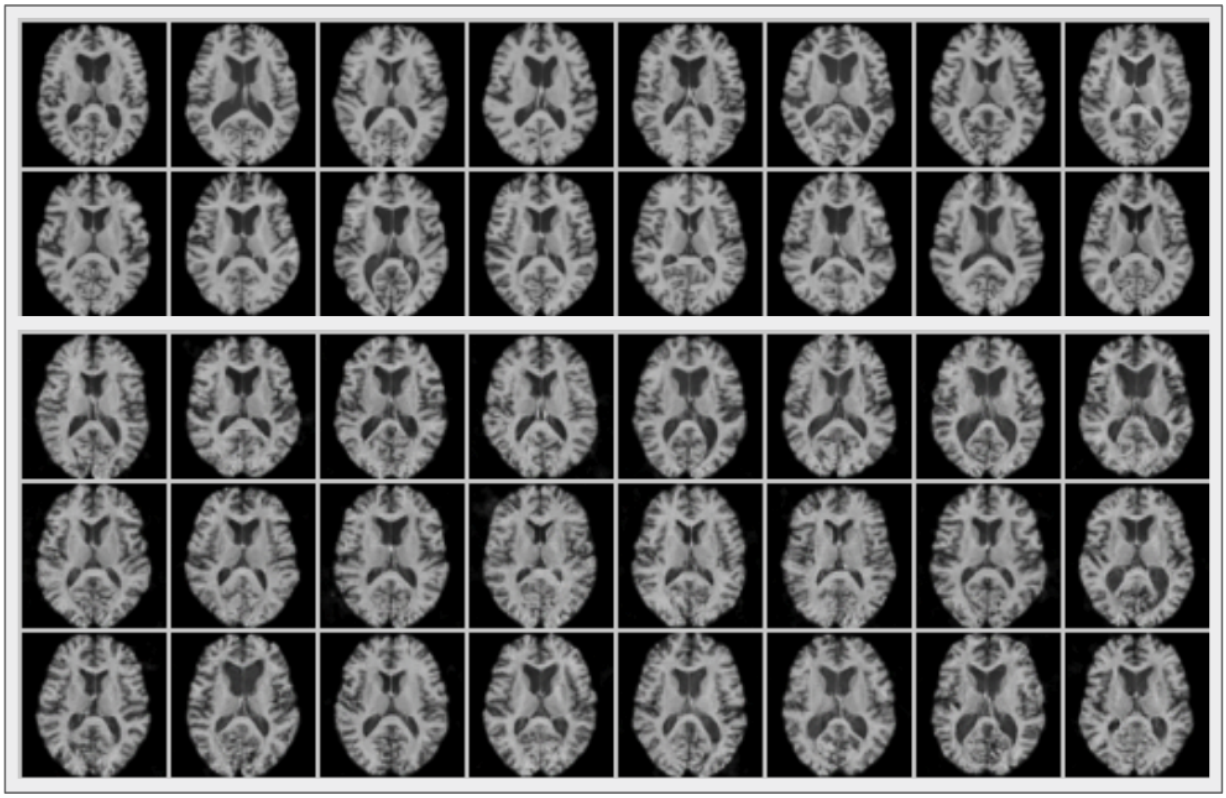


Figure 7.3: A random sample of synthetic images (bottom) of the ventricles produced by the GAN, with a selection of real images (top) for comparison.

also produced. Whilst encoding and reconstruction is only performed once in the following experiments, a lack of image degradation or "drift" after multiple applications is important as small errors become amplified. It is clear that the iterative approach yields more accurate reconstructions, even after multiple cycles. This improvement was considered to be worth the significantly longer encoding time and was therefore used in the following experiments.

7.3.3 Impact of re-weighting

The advantages of re-weighting can be seen in Figure 7.5 which shows a number of examples of cases with high levels of atrophy or other abnormalities which are accurately reconstructed using WGAN+RW, but which result in errors without re-weighting.

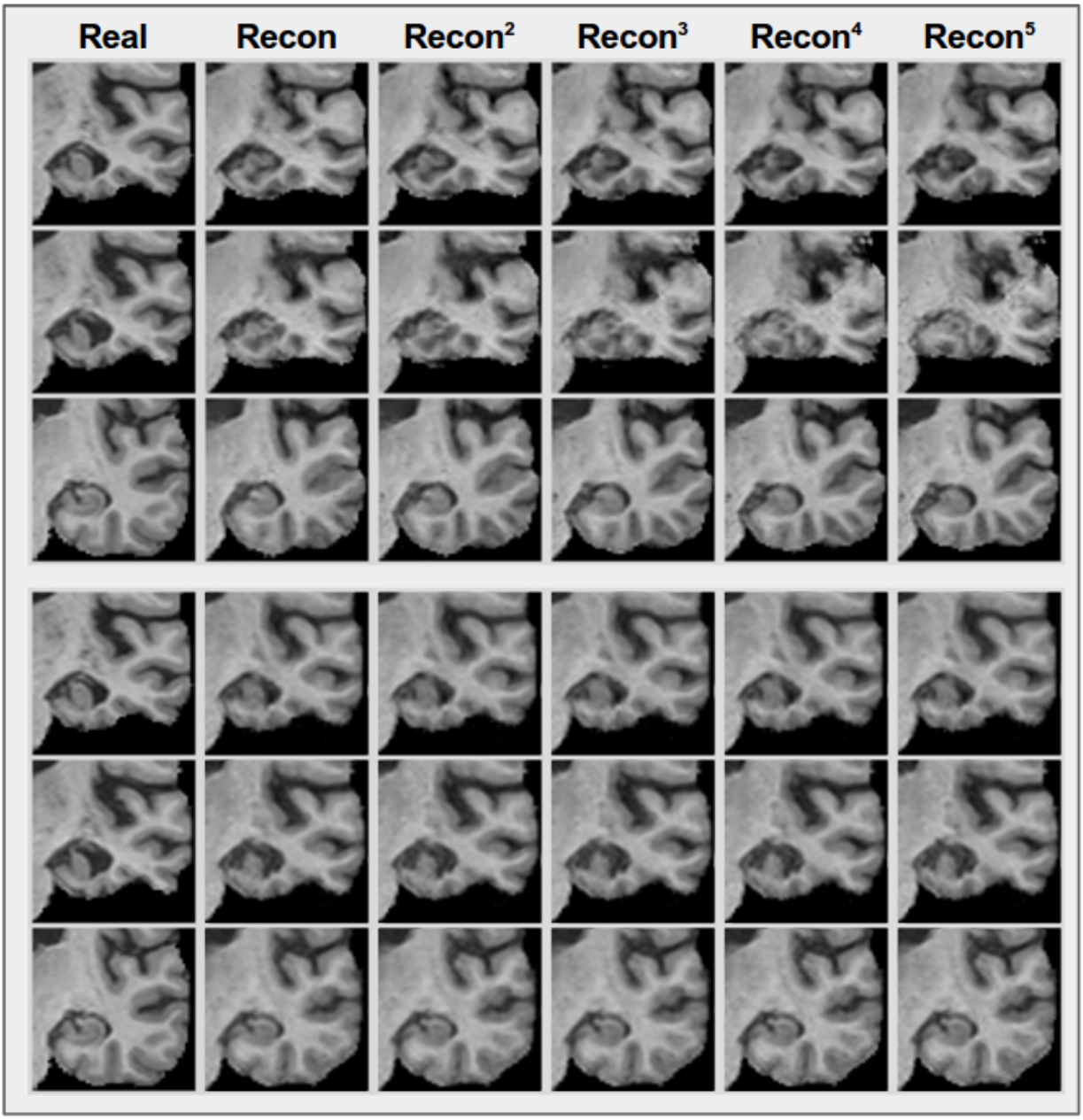


Figure 7.4: Examples reconstructions of three images using the two approaches. Top panel: retrained discriminator network. Bottom panel: gradient descent over **z**. **Recon** shows the reconstructed images according to the encoding calculated by each method. **Recon^{n}** shows the reconstructed images after **n** cycles of encoding and reconstruction. The first two images come from the same subject at different time points. The subtle differences in atrophy level are preserved using gradient descent over **z**, but lost using the retrained network.

7.3.4 Isolating the features of AD

Figure 7.6 shows the differences in average encoding between CN subjects and AD and MCI subjects for the hippocampal images. By examining the components which are statistically significantly (two-sample unpaired t-test, Bonferroni corrected for multiple comparisons) different

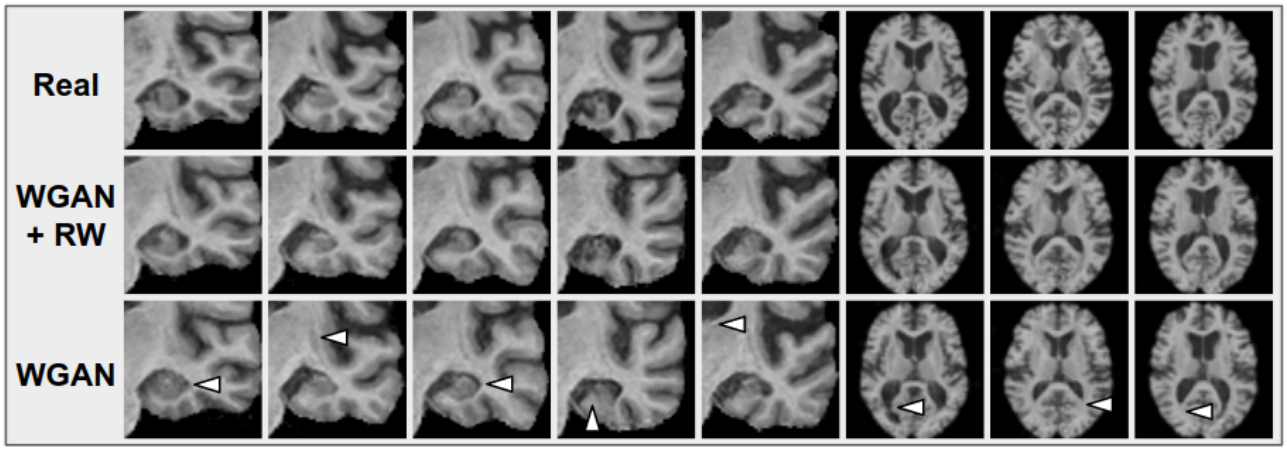


Figure 7.5: Selection of images showing cases where WGAN+RW leads to better reconstructions. Arrows indicate inaccurately reconstructed regions, each associated with a high level of atrophy or other abnormality.

between CN and AD subjects we can see which components are responsible for the differences seen between CN and AD images, and in which direction they move as images start to display stronger features of AD. As expected, the differences in these components between CN and MCI are in the same direction, but smaller, showing how the features progress as subjects move across the spectrum from CN, through MCI, to AD.

7.3.5 Adding and removing AD

Figure 7.7 shows the effect of adding multiples of the isolated features of AD in both locations. In both cases the subjects' general anatomy remains unchanged, however, the common changes seen in AD begin to manifest themselves as higher amounts of AD is added: Enlarged ventricles and cortical and hippocampal atrophy.

In figure 7.8, the effects of both adding and removing the features of AD from and AD subject with longitudinal data can be seen. AD is added to the baseline image, simulating the progression of AD over two years, and removed from the 24-month image, reverting it back to its baseline state. For the purposes of this experiment, $\mathbf{z}_{AD}$ was scaled such that $\mathbf{G}(\mathbf{z}^* - \mathbf{z}_{AD})$ visually resembled the baseline image.

The predicted images show the same effects as those seen in real AD patients: an increase

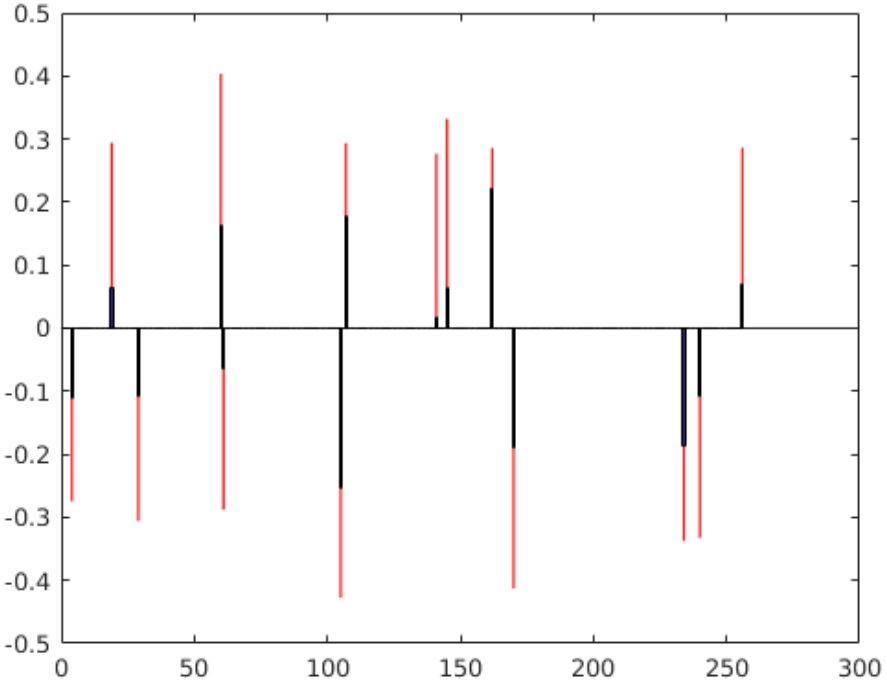


Figure 7.6: Visualisation of the average differences between the latent encoding of different subject groups in each location. Red: AD and CN. Blue: MCI and CN. Only elements with significant (two-sample unpaired t-test, $p < 0.05/256$) differences between AD and CN are shown.

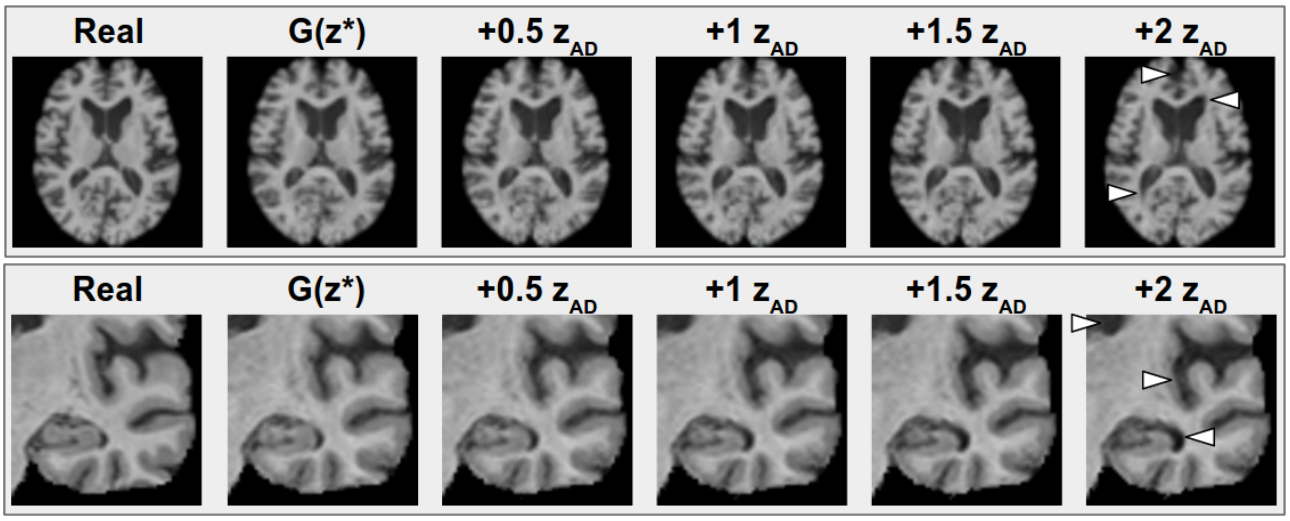


Figure 7.7: Progression showing the optimal reconstruction of a real image followed by reconstructions with multiples of $\mathbf{z}_{AD}$ added. Note the increasing presence of AD associated features. Top: Enlarged ventricles and cortical atrophy. Bottom: Enlarged ventricles, hippocampal atrophy and enlarged sulci.

in ventricle size over time. However, we can also see some of the limitations of the proposed method. The lack of resolution in the reconstructed images can mask the subtler changes and

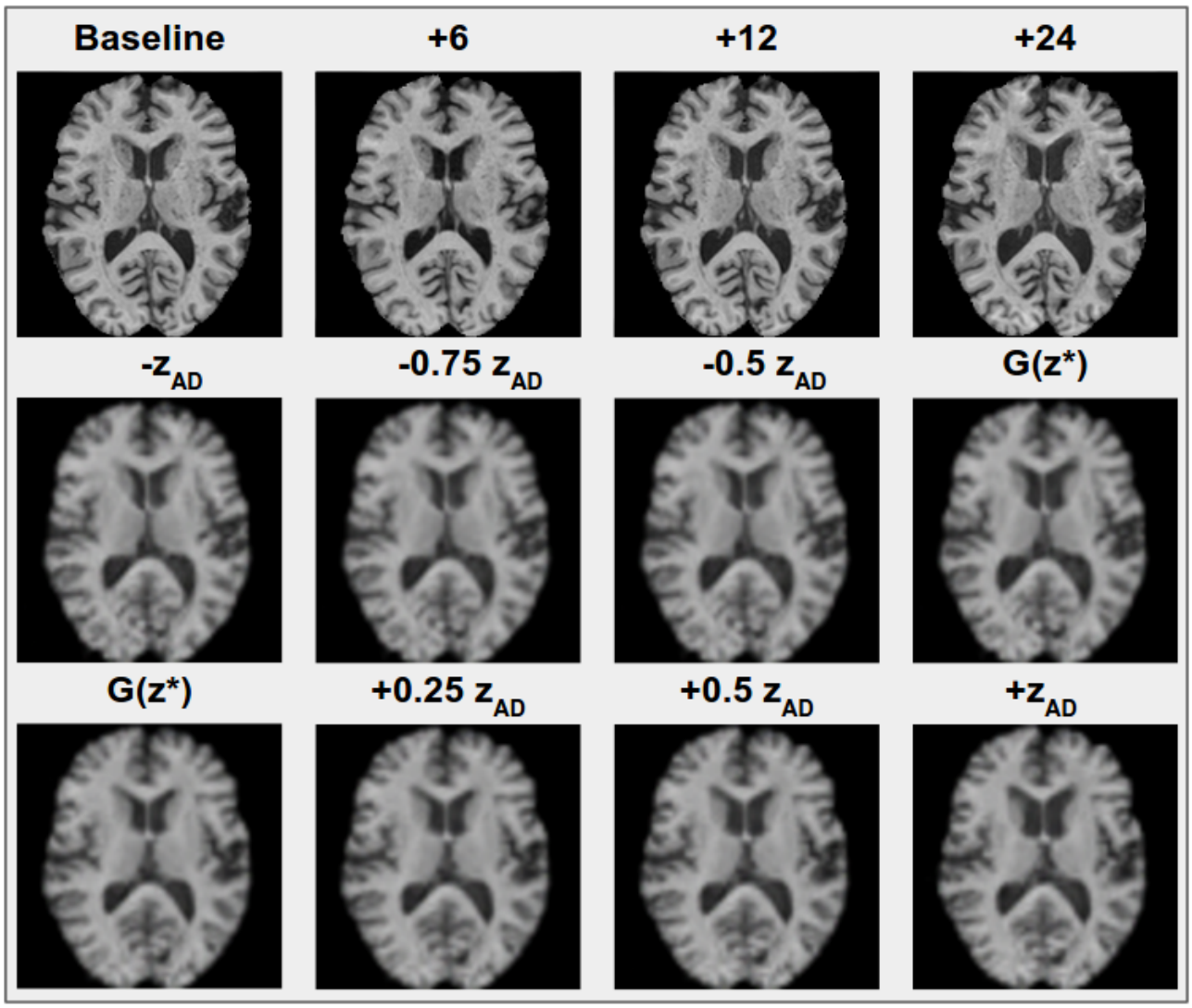


Figure 7.8: Images showing the addition and subtraction of AD features using a subject with temporal data. The top row shows the real images at full resolution for a subject with AD at baseline, 6 months, 12 months and 24 months. The second row shows the reconstruction of the 24-month image with predicted images for each other time point found by subtracting multiples of $\mathbf{z}_{AD}$. The bottom row shows the reconstruction of the baseline image with predicted images for each other time point found by adding multiples of $\mathbf{z}_{AD}$.

make accurate comparisons difficult. Also, the true atrophy progression is slightly asymmetric, with the effects being more prominent in the right hemisphere. This difference is respected when moving backwards from 24 months to baseline (it can't remove atrophy which isn't there) but is missed when moving forwards from the baseline. The predicted 24-month image displays more symmetric atrophy than seen in the real progression.

7.4 Future Work

Several significant assumptions were made in this work, a) that progression from healthy to AD is a linear process over time, b) atrophy patterns are symmetric, and c) morphological changes are the same across all patients with AD. Whilst these have been sufficient to demonstrate that the proposed method is able to make reasonable predictions, it is known that these assumptions do not hold (a [Jack et al., 1999], b&c [Rabinovici et al., 2008]). There are also many potential paths along which a healthy looking brain can progress to a more pathological appearing brain (for example, symmetrically or not), which means that progression is difficult to predict. However, given an true progression observed in a patient, it is a simpler task to revert the image back to a healthier state. Future work will involve creating a subject-specific latent AD encoding by incorporating clinical variables and subject anatomy to try and predict the mode and speed of progression most likely for that subject.

Previous work [Suk et al., 2015] has shown that compression of an image to latent space using stacked auto-encoders can lead to a powerful feature representation which can be used to classify patients across the CN-MCI-AD spectrum. Our approach leads to a similar compression which could also form the basis of a classifier. Figure 7.6 demonstrates this potential by showing that there are features which vary significantly between CN and AD, which could be used as a foundation for a classifier.

Our proposed method relies upon the GAN derived latent space facilitating image arithmetic. While this has been repeatedly demonstrated empirically [Radford et al., 2015, White, 2016], it is not clear whether this assumption always holds, and there is nothing in the GAN training procedure which guarantees this property. It is clear that, when pushed to the extremes, image arithmetic will fail. For example, continuously adding an AD vector eventually leads to no further change as we attempt to extrapolate beyond the training data distribution. Additionally, behaviour in these regions becomes unpredictable as the input vector will no longer be likely under the random z distribution on which the space was learned. The assumption of linearity is also not guaranteed, with linear changes in latent space not necessarily leading to linear changes in image space. However, despite the lack of any mathematical guarantees, our experiments

have shown that these assumptions do hold locally, if not globally, and provided changes are small and the resulting vector is still likely under the random z distribution, the results appear reasonable. Exploring the impact of these assumptions, and assessing under what conditions they can be held, will be a subject of future work.

The optimal size of latent encoding $\mathbf{z}$ could also be further investigated. Preliminary experiments suggested 256 provided a balance between the speed of convergence and the generation of visually realistic images. However, a deeper investigation into this, whilst extremely time-consuming, may yield better results. In addition, while the choice of using gradient descent over z instead a third network to map from image to latent space was well motivated given the results of the two methods, it may be the case that a third network could be the better option in other applications, particularly if time was a major consideration. Future work will involve investigating this further to establish under what conditions each approach is superior.

Finally, the proposed method's main limitation is the relatively small image size of 64x64 pixels which can be analysed. Whilst GAN formulations working at 128x128 [Berthelot et al., 2017] and 256x256 [Zhang et al., 2017] pixels have been proposed, these are often associated with long training times and large amounts of training data. 3-dimensional (3D) GANs are currently limited to binary objects occupying relatively small regions in image space [Wu et al., 2016], therefore containing much less complexity than MR images. The expansion of fast training GANs into higher image sizes, and ideally into 3D, would improve the power of the proposed method substantially.

7.5 Conclusion

We have demonstrated that the features corresponding to AD progression can be isolated using a GAN derived latent feature space. This was used to show that an accurate subject specific forecast for a typical AD progression can be found by reducing the subject's MR image to a latent feature space and adding the isolated latent encoding corresponding to the features of AD, before reconstructing the image from the new latent encoding. To aid in this, we have

also presented a modification to the WGAN training procedure which allows for more focus to be placed on the rare and extreme examples in the training data. We have shown that the proposed method produces images consistent with AD progression while maintaining each subject's unique anatomy and that this predicted progression corresponds well to that observed through longitudinal examination.

We hope that this work can lead to a better understanding of the pattern of AD progression, and provide a useful tool to both assess a patient's disease trajectory and monitor their response to treatment. Future work will involve calculating latent AD features which are tailored to the subject by taking into account current disease state and clinical scores. Extending GANs to higher image sizes and 3D MR images would also allow for more detailed predictions to be made.

Part A of this chapter has shown one application of the proposed method. In the next part we extend this work using an alternative architecture and investigate the ability of the proposed method to build more complex models and predictions based on age, disease state and genotype.

Chapter 8

Modelling the Progression of Alzheimer's Disease in MRI Using Generative Adversarial Networks - Part B

8.1 Introduction

In Part B of this chapter, we further develop the ideas presented in Part A. One of the limitations discussed previously is the relatively small region which can be investigated using the WGAN framework due to training instability at higher image sizes. During the time since the work described in Part A was carried out, alternative GAN formulations have been proposed, one of which, the Progressive Growing of GANs (PGGAN) [Karras et al., 2017], allows for significantly larger images to be generated. We, therefore, re-implement the methods from Part A of this chapter using PGGAN to further investigate their potential.

Early experiments demonstrated a reduced ability to accurately encode a real image using PGGAN. The potential reasons for this are discussed later. Because of this, the work in

this part moves away from subject-specific progression prediction, towards the visualisation of population-level changes throughout the stages of AD. Mentioned briefly in Part A of this chapter, this is another potentially useful application of the proposed method. In Part A we showed how the expected changes associated with AD can be imposed on an image. By applying this approach to a large number of subjects, it is possible to aggregate these changes across a population and visualise them. The differences in these average changes between different populations can then lead to potentially valuable insights. For example, in a drug trial, the effects of treatment can be visualised by comparing the average predicted changes between a treated and untreated group. While we maintain our focus on AD, this approach can be applied in any domain to observe the effect of any clinical variable on image data.

8.2 Materials and methods

The Alzheimers Disease Neuroimaging Initiative (ADNI) dataset used previously in Part A is used again, with a focus on the axial view which had to be downsampled to a 3mm resolution in the previous experiments, and can now be analysed at 1mm resolution.

A PGGAN is trained on the collection of single axial slices at full 1mm isotropic resolution, producing images of size 256-by-256. The PGGAN architecture and training procedure is unchanged from that described in [Karras et al., 2017]. The method from Part A for encoding a real image $\mathbf{I}$ to optimal latent vector $\mathbf{z^*}$ is modified to include an adversarial term. Whereas previously we aimed to minimise $|(\mathbf{G}(\mathbf{z^*}) - \mathbf{I})|^2$, this was found to lead to blurred images when using a PGGAN.

Introducing a discriminator loss from the already trained discriminator, as in [Yeh et al., 2016], was found to be insufficient to steer the search away from these blurry regions of the learned manifold. Instead, we choose to further train the discriminator by repeatedly showing it the real image we wish to encode. In this way, the discriminator no longer learns to separate real from generated images but instead aims to learn the features of the single target image, allowing it to guide the generator towards this image. We choose to reuse the discriminator in this way

as it is already a well-trained feature extractor for images of this type, and therefore requires little additional training.

This training is performed in parallel with the gradient descent procedure to find **z*****. We also initialised the search by finding the closest image as measured by Euclidean distance from a set of 10000 generated synthetic images previously produced by the generator, and used the corresponding latent vector $\hat{\mathbf{z}}$ to act as a starting point for the search. This ensures that the iterative procedure to find **z*** does not get stuck in a local minimum, far away from the optimal answer.

The final algorithm to encode image **I** to optimal latent vector **z*** is therefore shown in Algorithm 2.

Given image **I**; starting point $\hat{\mathbf{z}}$; generator G; discriminator D; loss weighting λ and number of iterations k
Set $\mathbf{z} = \hat{\mathbf{z}}$
for *i=1..k* **do**
  Compute $E^G = \mathrm{D}(\mathrm{G}(\mathbf{z}))$
  Compute $E^I = \mathrm{D}(\mathbf{I})$
  Update parameters of D, θ_d, by ascending $\Delta_{\theta_d}[E^I - E^G]$
  Compute $E^G = \mathrm{D}(\mathrm{G}(\mathbf{z}))$
  Compute $E^L = |(\mathbf{G}(\mathbf{z}) - \mathbf{I})|^2$
  Update latent encoding, $\mathbf{z}$, by descending $\Delta_{\mathbf{z}}[E^G + \lambda E^L]$
end

Algorithm 2: Updated encoding procedure, introducing an adversarial loss from real image **I**.

8.2.1 Image quality

Figures 8.1 and 8.3 show un-curated generated images. The image quality appears qualitatively high, with high variation and little or no visible differences from real images, a sample of which can be seen in Figure 8.2.

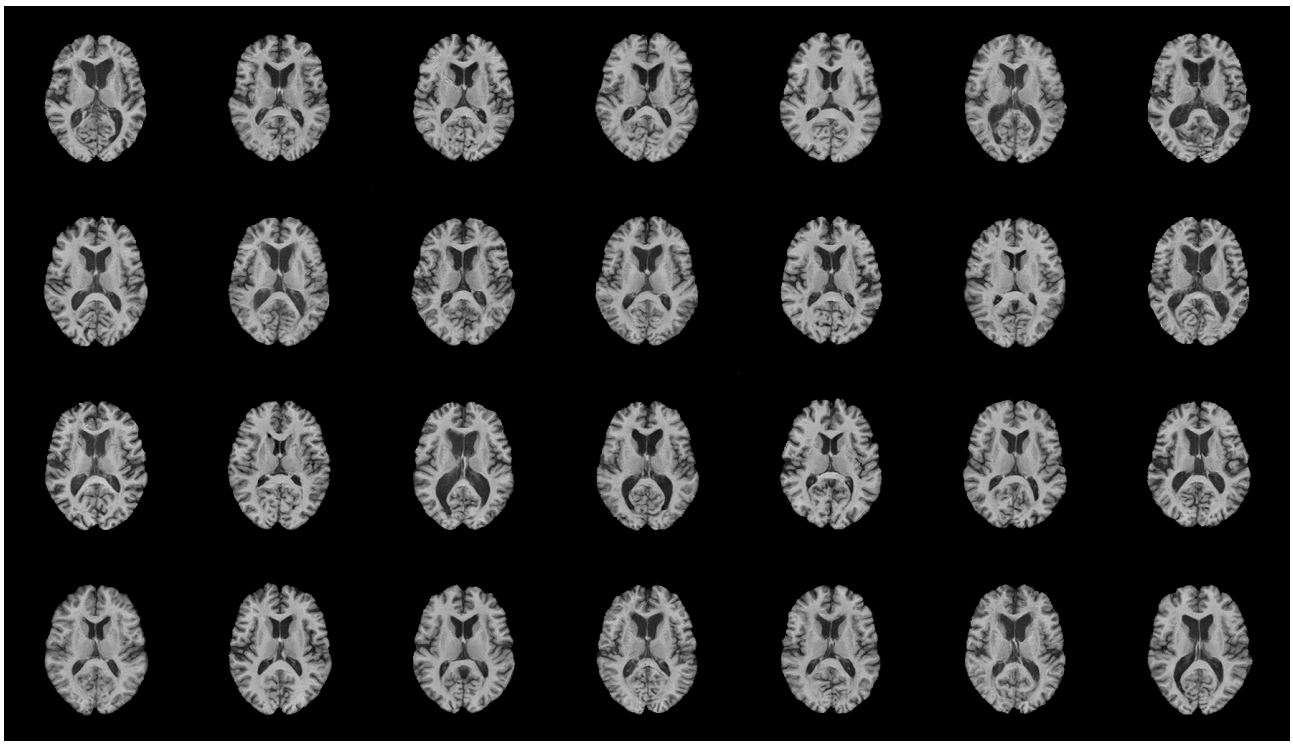


Figure 8.1: A random sample of synthetic images.

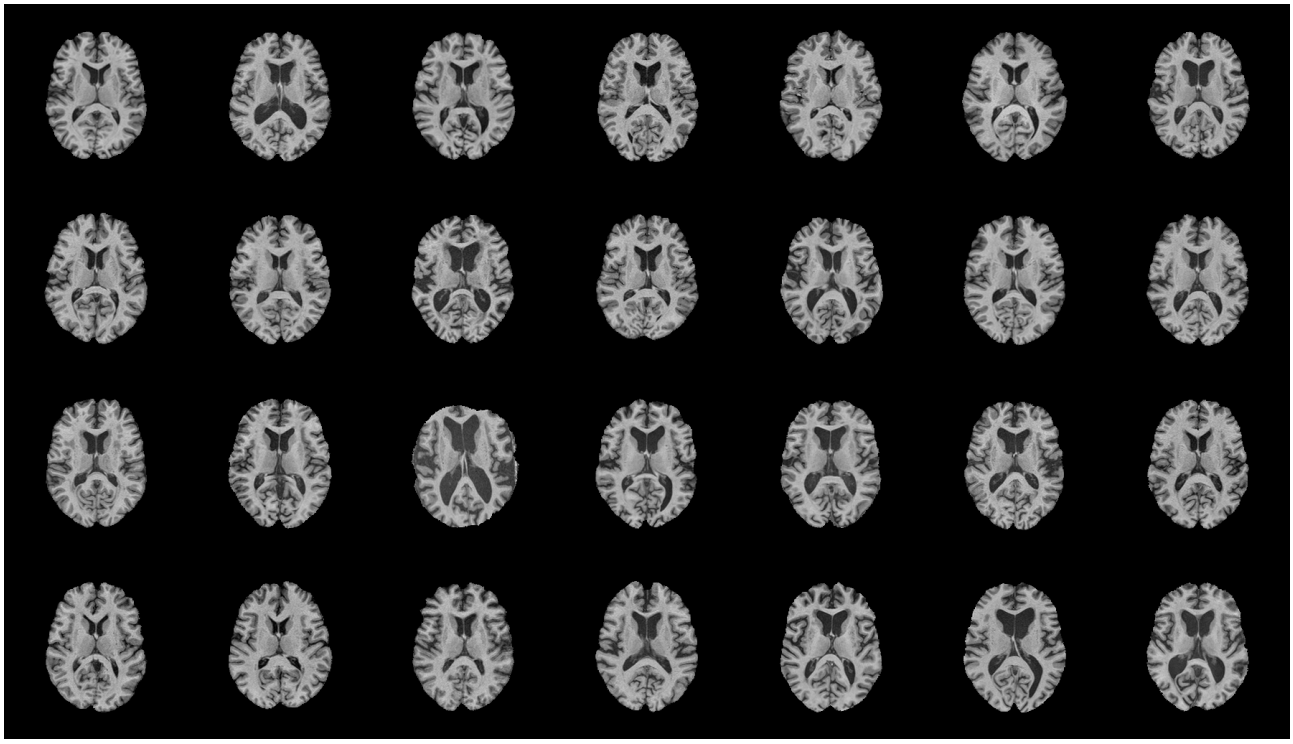


Figure 8.2: A random sample of real images.

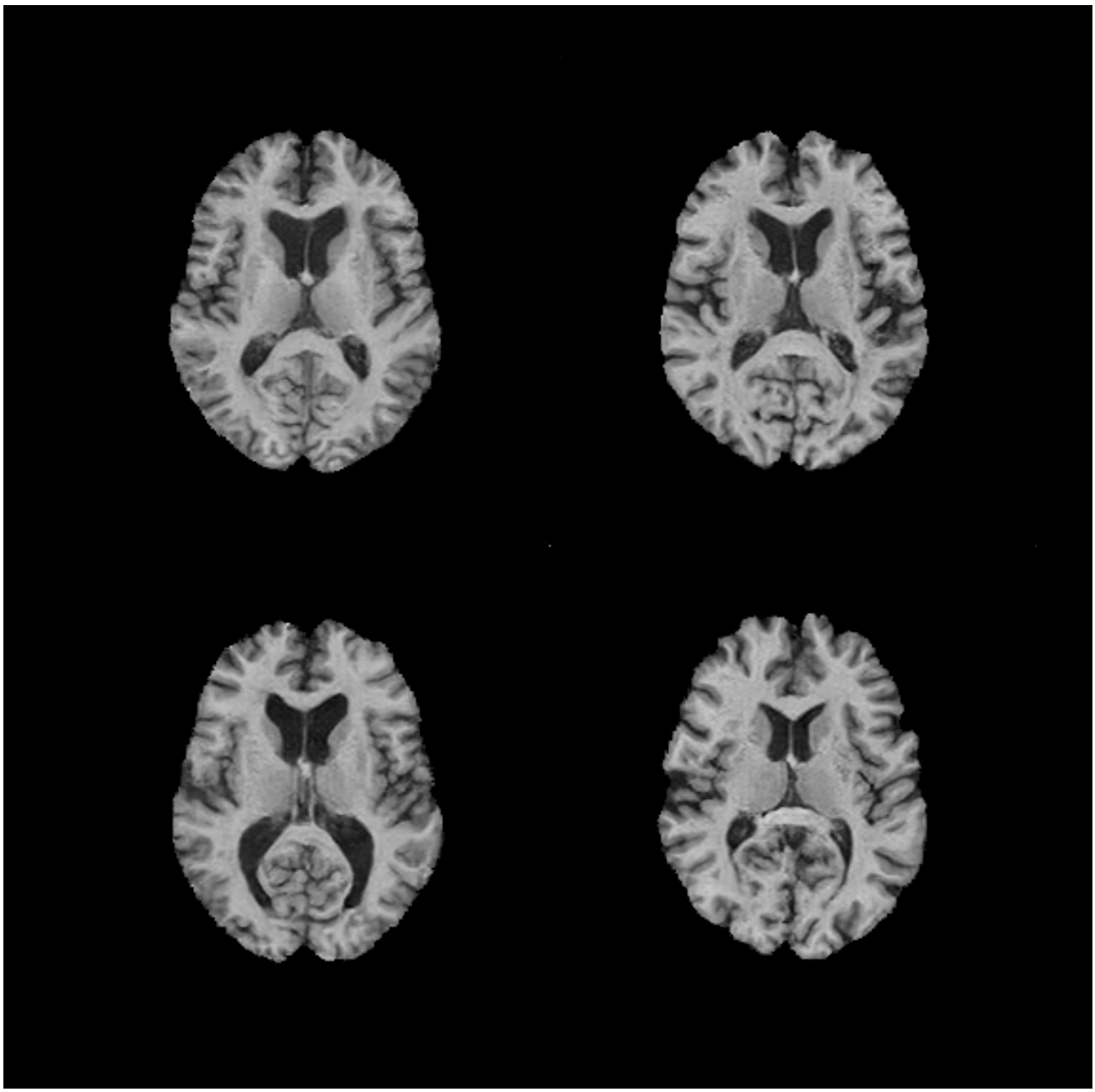


Figure 8.3: A random sample of synthetic images (full resolution).

8.2.2 Single image encoding

Prior to running any experiments, we investigate the accuracy to which a latent encoding can be generated for a real image using Algorithm 2 compared with the encoding procedure used in Part A. Figure 8.4 shows how the original procedure tends to generate blurry images. This comes as a result of there being small regions of the learned manifold which map to unrealistic blurry images. A visual analysis of random synthetic images suggests that less than 0.5% of the

manifold surface maps to these regions, however, they act as a sink point in almost all intensity based searches for an optimal encoding. We also found that the discriminator considered images sampled from this region as being real, despite their blurry appearance, hence incorporating a trained discriminator loss, as in [Yeh et al., 2016], proves unsuccessful. However, training an image specific discriminator to differentiate synthetic image from the single real image **I**, leads to a discriminator loss which proved effective at ensuring a realistic image. A sample of these can be seen in Figure 8.5.

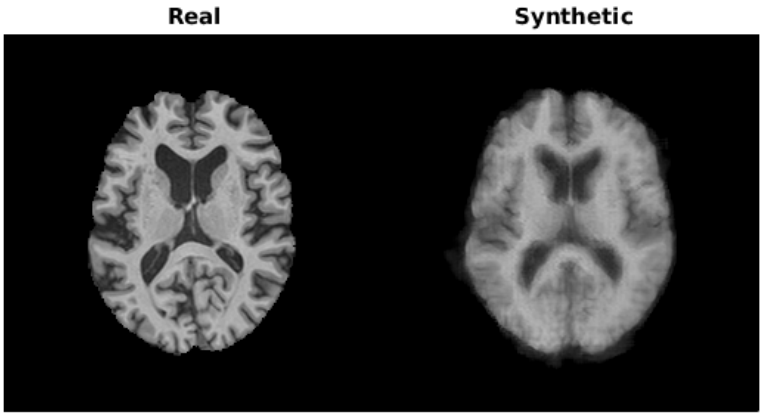


Figure 8.4: An encoded image using the method used previously in Part A.

Despite appearing visually similar, the reconstructed images appear to have more errors than was observed in Part A when using the WGAN on smaller images. There are several potential causes of this. Firstly, the higher image sizes mean a greater level of variation which must be matched in the generated images. This leads to a more complex search for $\mathbf{z}^*$ with more potential to end up in local minima. Another cause could be the learned manifold itself, while the majority of the generated images appear realistic, there is no guarantee that all possible images have a representative point on the manifold. There may be too much variation to be encoded in a 512 element latent vector. The learning procedure of the PGGAN may contribute to this. Once the higher resolution levels are reached, and the highest frequency information becomes visible to the GAN, all the available source of entropy could have already been used to produce variation in the lower frequency features. This could result in two behaviours. First, the GAN could reuse entropy previously used for a low-frequency feature, such as ventricle size,

for the purposes of describing high-frequency features, such as a particular region of cortical folding. This would result in the two features being coupled, with one not generated without the other, collapsing the manifold. Alternatively, the GAN could learn later layers which effectively perform a super-resolution procedure based upon the output of the previous layers. The manifold is learned at a low resolution, with the later layers simply adding detail on top of this. However, since it would only learn one high-resolution image from each point on the low-resolution manifold, many real images cannot be fully represented. The exact cause of this, as well as the source of the blurry regions, would involve an extensive investigation into the behaviour of the PGGAN architecture and training procedure. However, despite these limitations, we proceed with our analysis. Whilst we are unable to accurately generate a subject-specific image to produce a prediction of progression, as done in Part A of this chapter, by analysing a sufficiently large amount of data we can perform population-level analysis and visualise the changes on synthetic images.

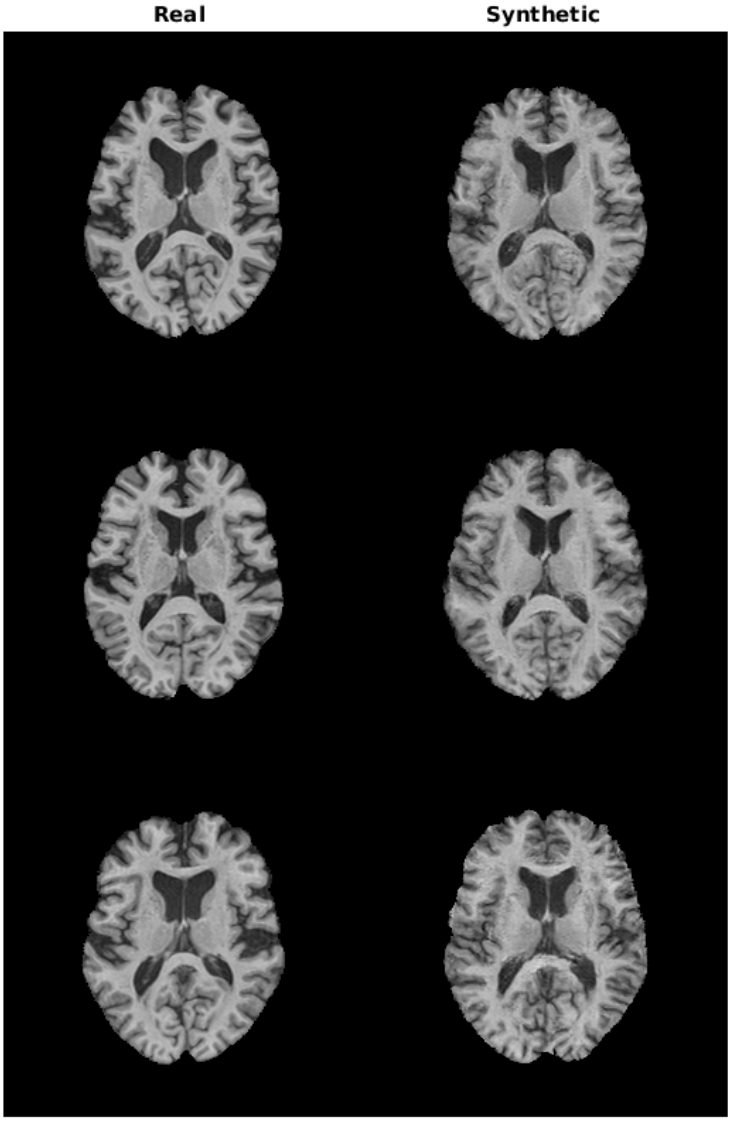


Figure 8.5: A set of encoded images using the method given in Algorithm 2.

8.3 Experiments

We run the following four experiments to investigate the ability of the proposed method to detect the changes present in AD. In the first two, we look to validate the proposed method by testing its ability to identify the expected changes in AD, as well as the differences between each disease state.

- The first experiment mirrors those performed earlier in Part A of this chapter, isolating the latent signature for AD and using this to add the features of AD to generated images.

- We then extend this further to other disease states, examining the differences between each pair of disease states, and averaging the predicted differences over the entire dataset to visualise the average expected changes.

- We next incorporate age into the model by fitting a non-parametric kernel regression model in latent space relating age to expected changes in each of the latent vector components. This allows us to compare the expected year-to-year changes, and, by averaging these, visualise the expected rate of change under each disease state.

- Having verified that the method can detect the more obvious changes, we proceed to a more challenging task and perform comparisons with a simpler method. We examine whether the proposed method can identify associations between clinical variables and anatomical changes by isolating the effect of the Apolipoprotein E (APOE) genotype on the expected rate of change of subjects with Late Mild Cognitive Impairment (LMCI) and AD. Finally, we implement an alternative model based on pixel intensities in image space, demonstrating the benefits of modelling in a latent space rather than image space.

8.3.1 Global differences in CN, Early Mild Cognitive Impairment (EMCI), LMCI and AD

Our first experiment reflects those performed in Part A of this chapter. We take the average computed optimal latent encodings of images of subjects which are healthy controls, and those with EMCI, LMCI, and AD. For this experiment we use only the baseline images, providing 182 CN, 175, EMCI, 158 LMCI and 137 AD images. Signatures describing the difference between the average encodings for each disease state ($\bar{\mathbf{z}}_{\{CN,EMCI,LMCI,AD\}}$) and each other disease state are computed, for example, $\mathbf{z}_{CN,AD} = \bar{\mathbf{z}}_{AD} - \bar{\mathbf{z}}_{CN}$. Figure 8.7 shows the effect of adding these signatures to random generated images.

We next add the appropriate signature to the latent encodings of all subjects belonging to each disease state to produce a predicted image under each other disease state. For example, for each CN subject, a predicted encoding for that subject under EMCI, LMCI and AD are generated by adding $\mathbf{z}_{CN,EMCI}$, $\mathbf{z}_{CN,LMCI}$, and $\mathbf{z}_{CN,AD}$ respectively. These sets of latent encodings are then passed through the generator and the resulting images are averaged. By examining the differences between these average images, we can visualise the expected changes in image intensity as groups progress from one disease state to the next. Figure 8.6 shows these expected differences.

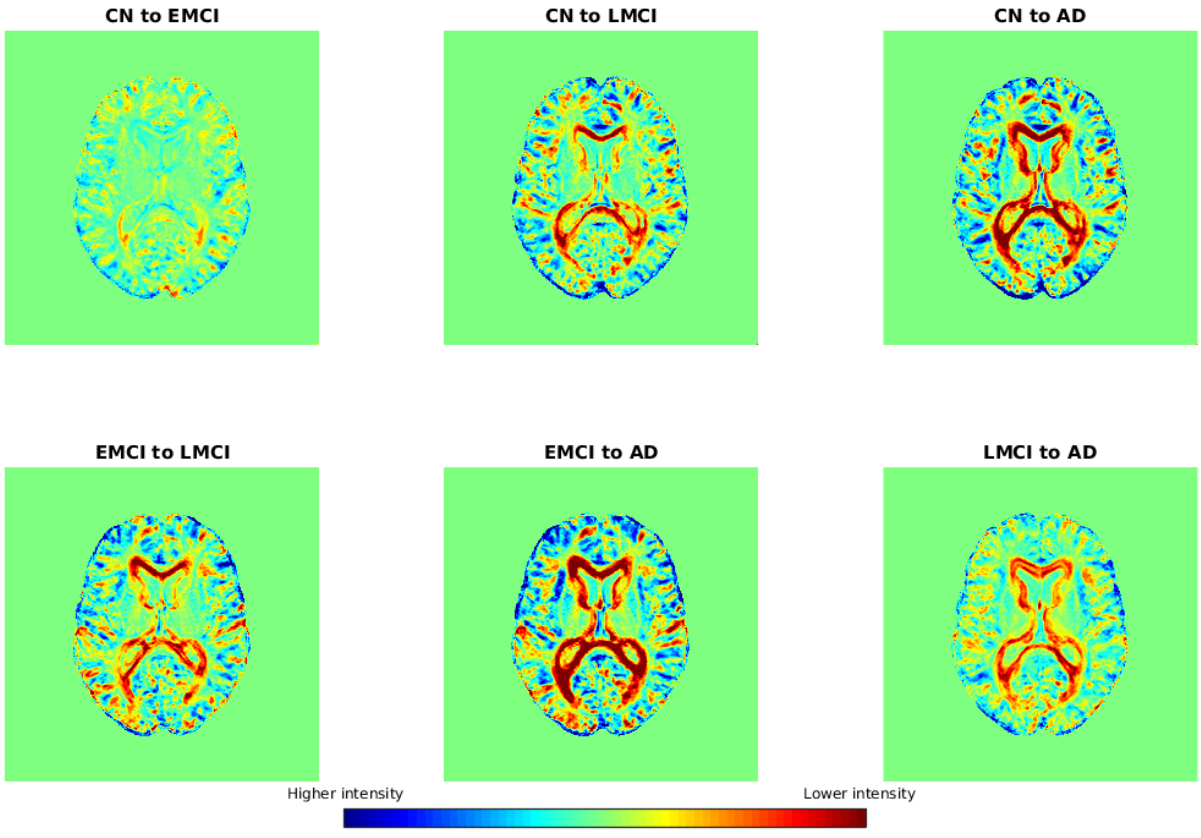


Figure 8.6: Average differences observed when turning a set of images corresponding to one disease state into another by adding the corresponding disease signature.

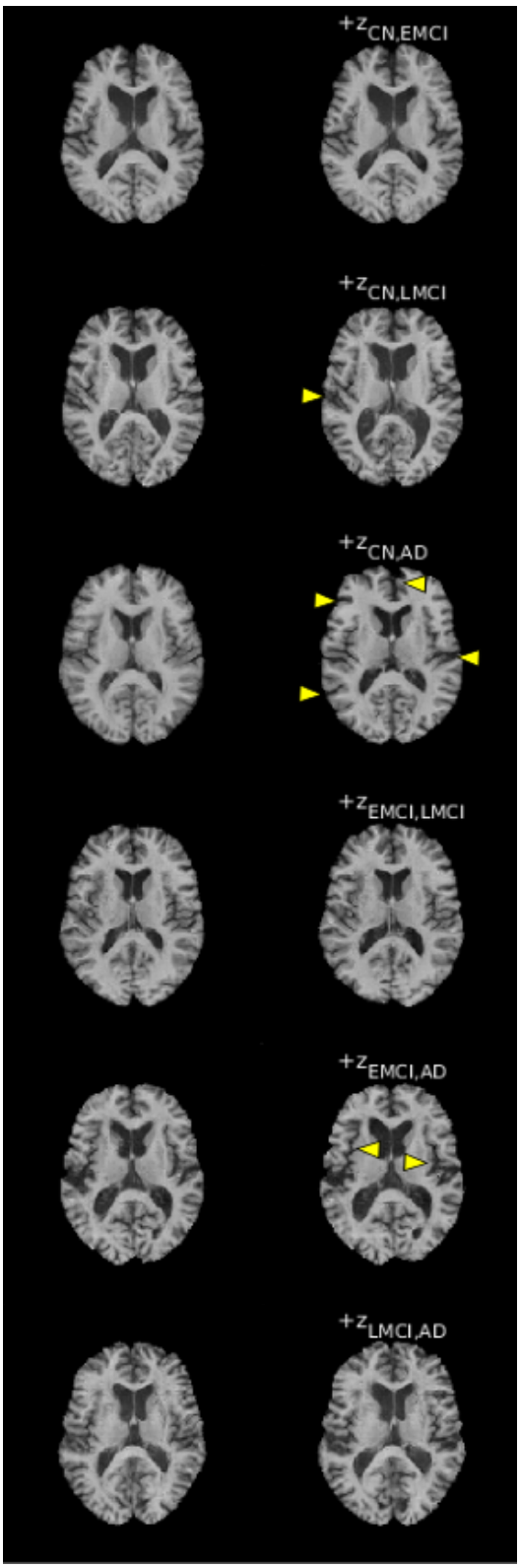


Figure 8.7: The effect of adding different disease signatures to random images. There is clear ventricular enlargement between all disease states except CN and EMCI which appears mostly unchanged. Arrows indicate some regions of increased cortical atrophy, which are more clearly seen when moving from CN or EMCI to LMCI or AD.

8.3.2 Predicted year-to-year tissue changes

In this experiment we look to utilise information about the age of the subjects to produce a model which will show us the expected year-to-year tissue changes in each disease state, and compare these to that of healthy ageing. A model is produced for each component of an image's latent encoding describing how it is expected to change as age and disease state varies. For each component, a non-parametric model $\mathbf{M}$ is fitted using kernel regression with a standard deviation of $h = 5$ years, relating age to the value of the that component in the computed optimal latent encodings for each disease state. An estimate for the value of latent component indexed at $x \in \{1..|\mathbf{z}|\}$ at age α for disease state S with associated set of latent vector components $\mathbf{Z}_{x,S}$ and corresponding ages $\boldsymbol{\alpha}$ is therefore given by:

$$\mathbf{M}_{x,S}(\alpha) = \frac{\sum_i (K((\boldsymbol{\alpha}_S(i))/h)\mathbf{Z}_{x,S}(i))}{\sum_i K((\alpha - \boldsymbol{\alpha}_S(i))/h)}, \tag{8.1}$$

$$K(p) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}p^2}. \tag{8.2}$$

Figure 8.8 shows an example of this. While all the data is used to compute the model, we restrict the domain to between the ages of 60 and 85, due to the scarcity of data points outside this range reducing the reliability of the model for these values. The computed models for four latent components for each of the disease states are shown in Figure 8.9. Using these models, a predicted latent encoding for each subject at each age between 60 and 85 is then produced. To predict component x at of latent encoding $\mathbf{z}$ at age α using model $\mathbf{M}$ for a subject with disease state S with an image taken at age α_0 with optimal encoding $\mathbf{z^*}$, the following equation is used:

$$\mathbf{z}_{x,\alpha} = \mathbf{z^*}_x + [\mathbf{M}_{x,S}(\alpha) - \mathbf{M}_{x,S}(\alpha_0)]. \tag{8.3}$$

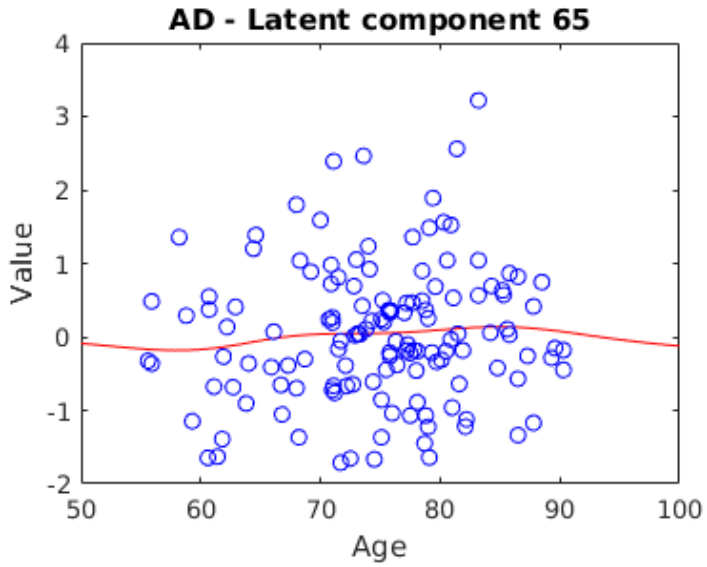


Figure 8.8: Example of using kernel regression learn a relationship between age and one latent component for subjects with AD. This shows how the expected value of a particular latent component varies for subjects at different ages. We can see that, within the truncated range of [60,85], the predicted value gradually rises from below to above 0.

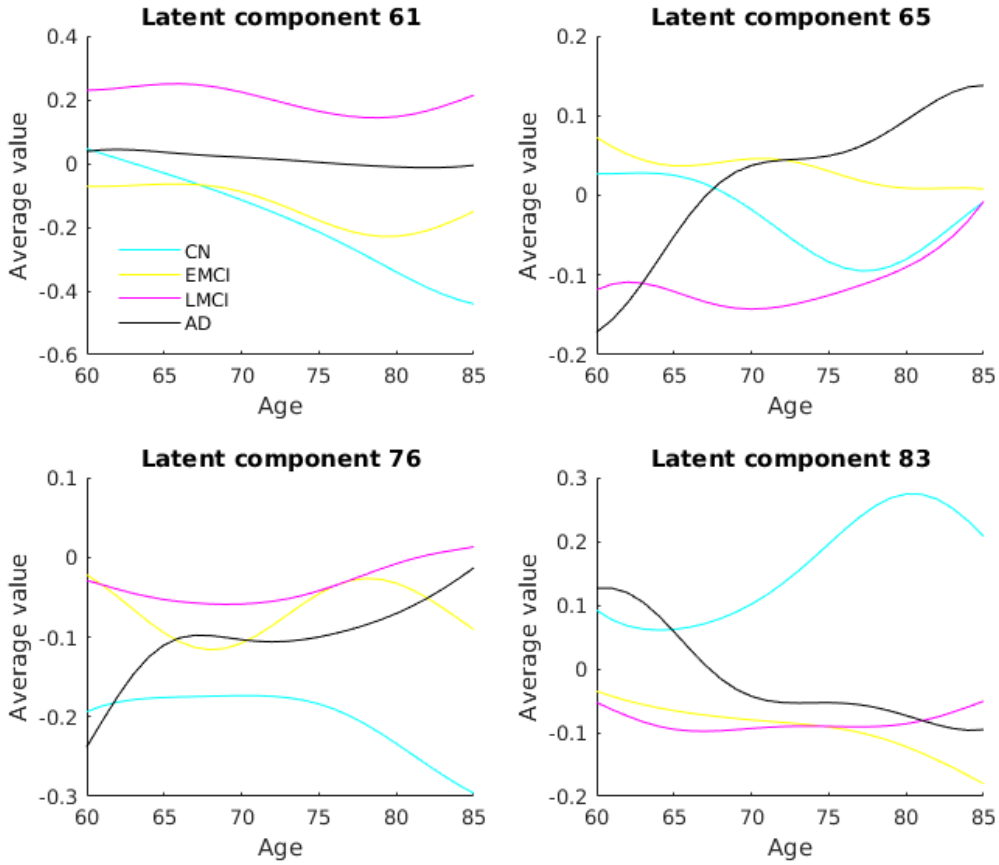


Figure 8.9: Models learned for four latent components for each disease state with their domains restricted to between 60 and 85.

Once computed, these latent vectors are passed through the generator to give a set of images describing a subject's predicted progression from 60 to 85. One of these for each disease state is shown in Figure 8.10. By taking the average over all subjects, an atlas is formed for each disease state at each age. Figure 8.11 shows the average difference in intensity between each such atlas and a baseline defined as the atlas computed for a 60-year-old CN subject.

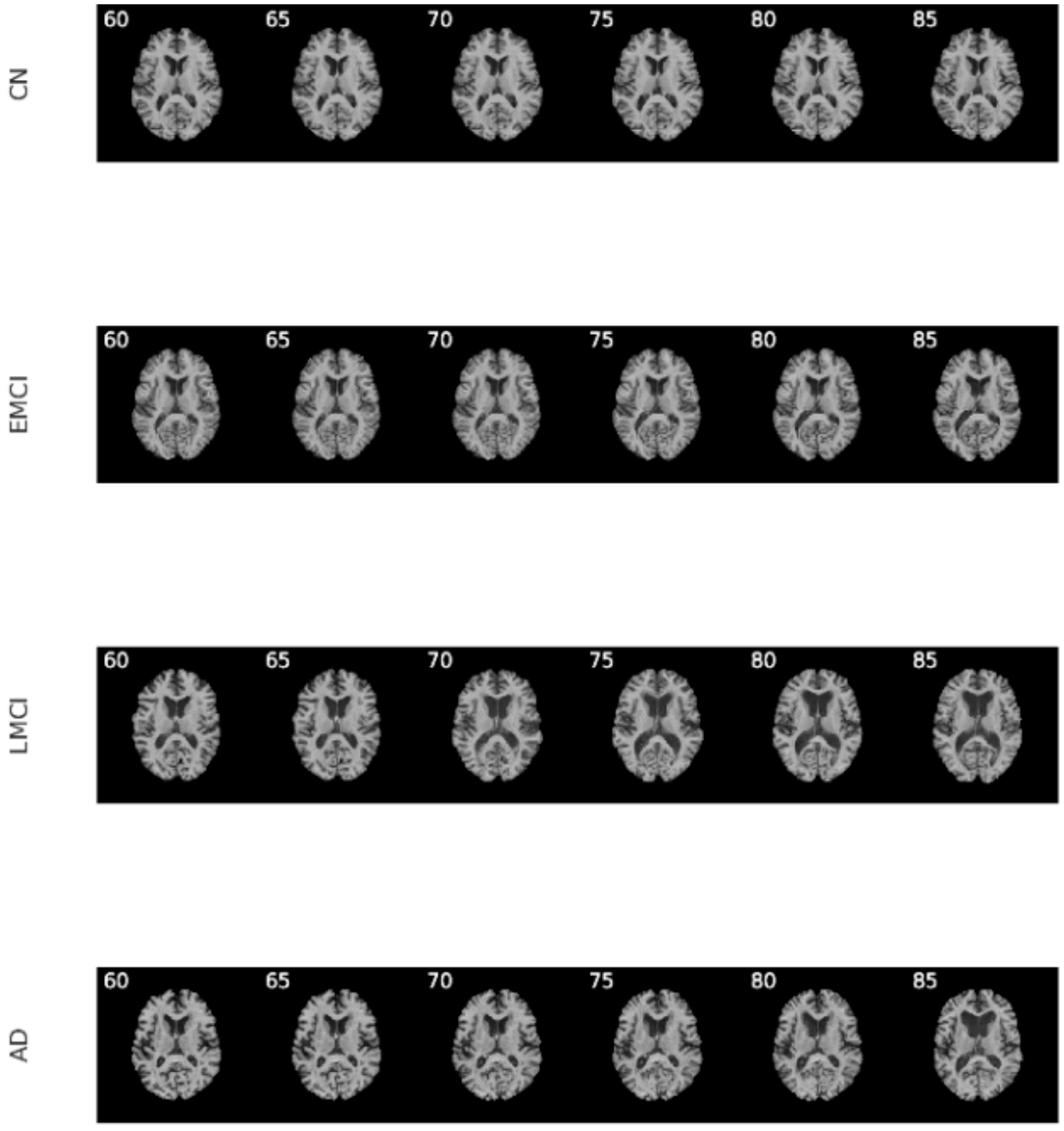


Figure 8.10: Predicted progression of a synthetic image between the ages of 60 and 85 for each disease state.

A rate of change is then found for each disease state by examining the difference between atlases from year to year. The average difference across each pair of consecutive years is computed, giving a single image for each disease state showing the predicted rates of change of image intensity, which can be seen in Figure 8.12.

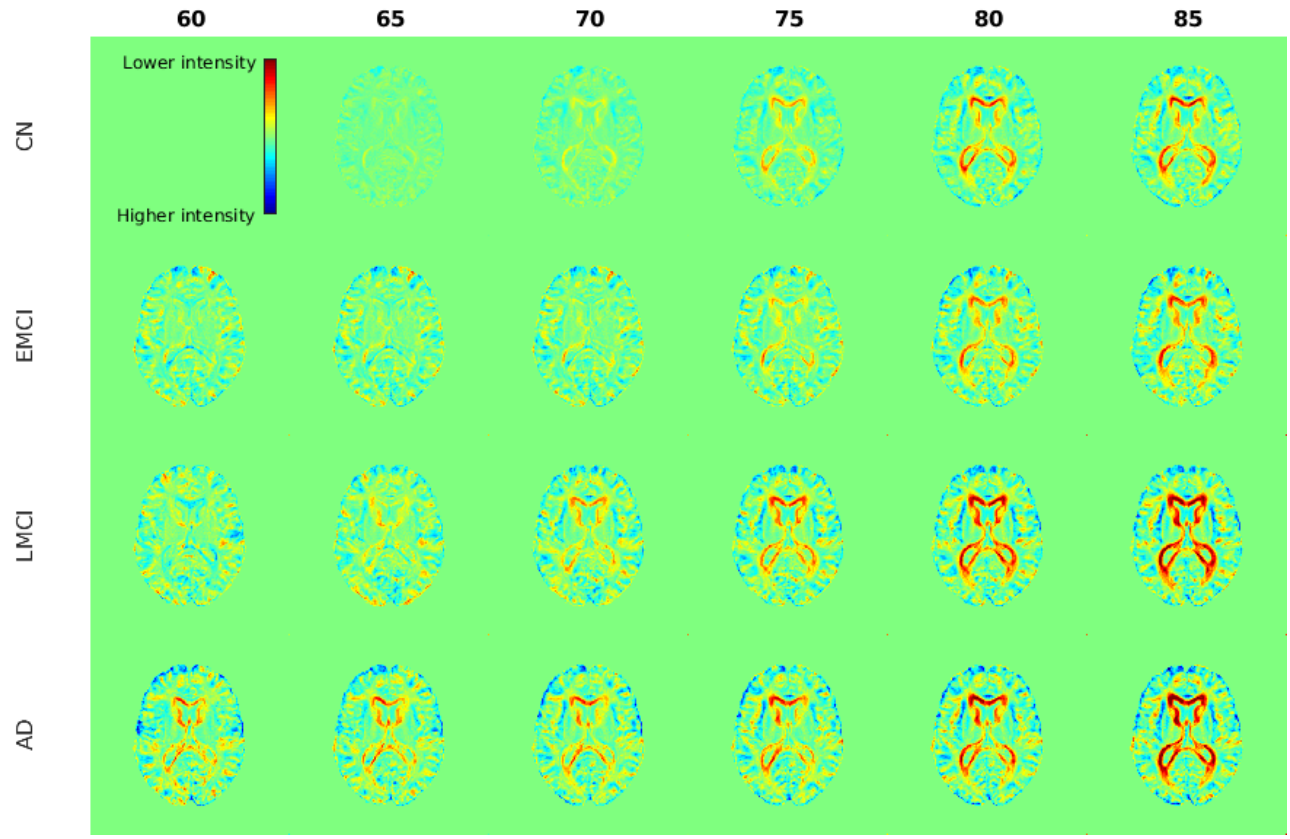


Figure 8.11: Differences in image intensity between atlases predicted for each disease state and age, and a baseline defined as the predicted atlas for a 60 year old CN subject.

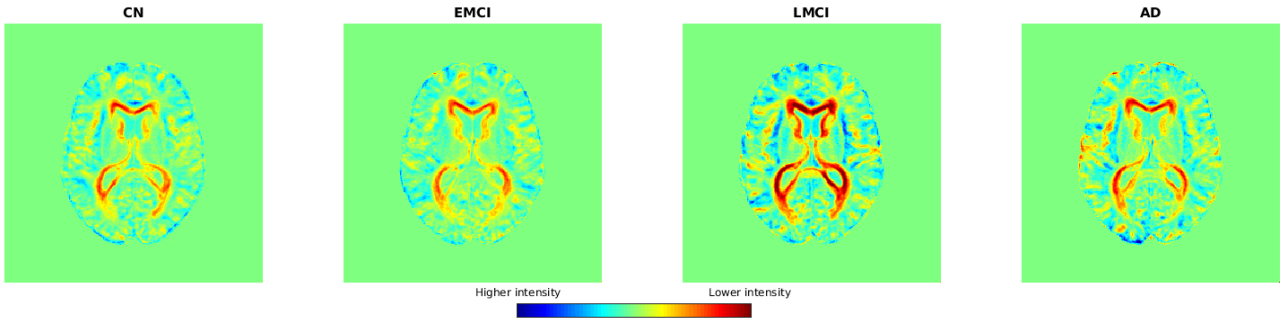


Figure 8.12: Average difference in predicted atlas intensities across each pair of consecutive years for each disease state.

8.3.3 Effects of apolipoprotein E in LMCI and AD

It has been shown that the APOE is associated with the risk of developing AD (see Table 2.1), with 1 or 2 copies of the $\epsilon 4$ allele being associated with a 300% and 800% increased risk of developing AD respectively. It has also been reported [Nestor et al., 2008] that having at least one copy of the $\epsilon 4$ allele increases the rate of ventricular enlargement, particularly in AD. As a test of the proposed method's ability to identify associations between clinical variables and image properties, we repeated the previous experiment with the cohort split into two groups, one containing the subjects with no copies of the $\epsilon 4$ allele ($\epsilon 4^-$), and the other containing subjects with one or more copies ($\epsilon 4^+$). To increase the sample size, the LMCI subjects were also used in addition to the AD subjects. This gives 113 $\epsilon 4^-$ cases and 182 $\epsilon 4^+$ cases. The difference between the calculated average rates of change can be seen in Figure 8.13. To examine the stability of the method and to get a measure of the error in the predictions we perform bootstrapping by repeating the experiment 50 times using a different subset of cases randomly sampled with replacement from each group for each run. To ensure no data imbalance, only 113 cases are randomly sampled from each group. The results averaged across all runs can be seen in Figure 8.14 along with a t-score map showing regions of significant difference.

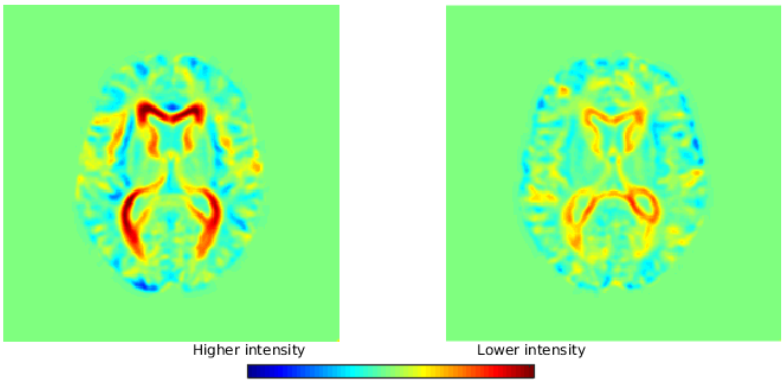


Figure 8.13: Average year-to-year change in image intensity for $\epsilon 4^+$ (left) and $\epsilon 4^-$ (right) across all subjects.

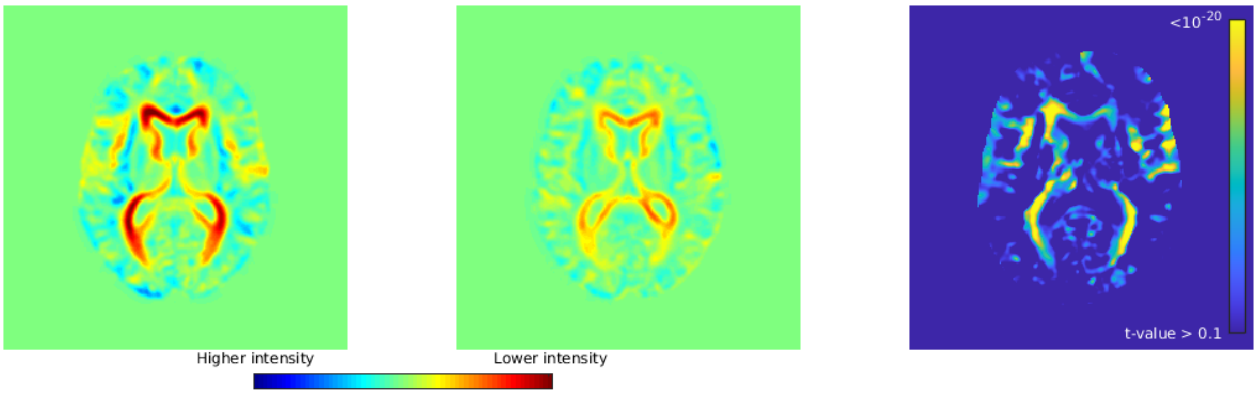


Figure 8.14: Average year-to-year change in image intensity for $\epsilon 4^+$ (left) and $\epsilon 4^-$ (middle) subjects, averaged over 50 runs, and t-value (right) map indicating regions of a significantly higher rate of intensity change in $\epsilon 4^+$ subjects.

8.3.4 Comparison to direct image analysis

In this section, we repeat some of the above experiments in image space in order to highlight the similarities and differences to the proposed method.

The proposed method learns a model for age in the GAN derived latent space of the image data. By fitting the same kernel regression model to each pixel in image space, a similar model can be formed relating pixel intensity to age across different disease states. While this model cannot be used to make subject specific predictions, it does predict average year-to-year changes in pixel intensity. Specifically, the expected intensity difference δ from age α_0 to α_1 at pixel x for disease state S can be calculated from image intensities $\mathbf{I}_{x,S}$ and corresponding ages $\boldsymbol{\alpha}_S$ as:

$$\delta_{x,S}(\alpha_0, \alpha_1) = \mathbf{M}_{x,S}(\alpha_1) - \mathbf{M}_{x,S}(\alpha_0), \tag{8.4}$$

$$\mathbf{M}_{x,S}(\alpha) = \frac{\sum_i (K((\boldsymbol{\alpha}_S(i))/h)\mathbf{I}_{x,S}(i))}{\sum_i K((\alpha - \boldsymbol{\alpha}_S(i))/h)}, K(p) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}p^2}, \tag{8.5}$$

where $k = 5$ to reflect the previous experiments. Using Equation 8.4, the average year-to-year change in pixel intensity between the ages of 60 and 85 is then calculated as:

$$\bar{\delta}_{x,S} = \frac{1}{25} \sum_{\alpha=60}^{84} \delta_{x,S}(\alpha, \alpha + 1). \tag{8.6}$$

Figure 8.15 shows these changes for each disease state, while Figure 8.16 shows the same calculated over $\epsilon4^+$ and $\epsilon4^-$ LMCI/AD subjects. For the latter, the same bootstrapping procedure as used previously is employed.

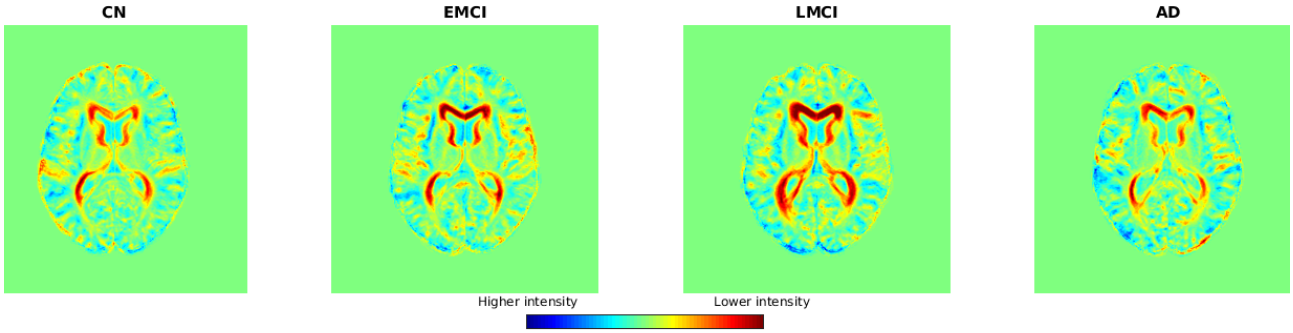


Figure 8.15: Average year-to-year changes for each disease state calculated in image space.

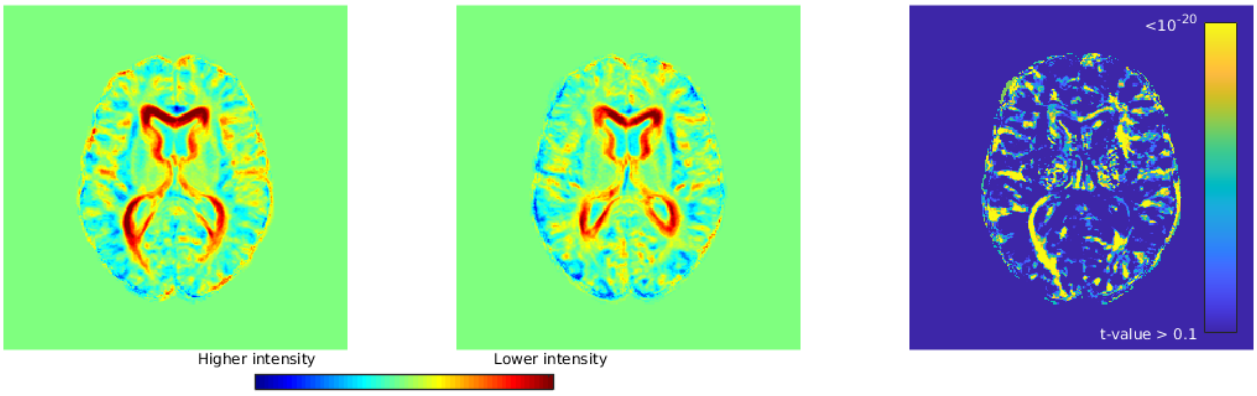


Figure 8.16: Average year-to-year change in image intensity calculated in image space for $\epsilon4^+$ (left) and $\epsilon4^-$ (middle) subjects, averaged over 50 runs, and t-value (right) map indicating regions of a significantly higher rate of intensity change in $\epsilon4^+$ subjects.

8.4 Discussion

Global differences in CN, EMCI, EMCI and AD

Figure 8.7 shows the expected changes as a synthetic image is changed from one disease state to another. As expected, we see the biggest change when going from CN to LMCI and AD, while we see very little change from CN to EMCI. The changes observed from EMCI to LMCI and LMCI to AD are comparatively small in magnitude, but reflect those expected as the disease progresses - particularly the enlargement of the lateral ventricles. When making larger changes, from CN to LMCI or AD, or EMCI to AD, we can also observe a slight widening of the Sylvian fissure, and in the case of CN to AD, atrophy at various points around the cortex become apparent. When viewing the average changes across the dataset in Figure 8.6, we see the same patterns. The primary differences between disease states appear around the ventricles, as well as some evidence of cortical atrophy between CN/EMCI and LMCI/AD. Again, these findings agree with the expected changes throughout the progression of AD, with ventricles and sulci appearing enlarged as surrounding tissue atrophies.

8.4.1 Predicted year-to-year tissue changes

Figure 8.10 shows the predicted changes under each disease state between the ages of 60 and 85 projected onto a random image. Under normal ageing, we see little if any change, however under EMCI we begin to see the ventricles appearing slightly enlarged, and, under LMCI and AD, we see significant ventricular enlargement. When we look at the average differences from a theoretical atlas for a healthy 60 year old in Figure 8.11, we see a clear pattern of ventricular enlargement across each disease state, being much stronger in LMCI and AD than CN and EMCI. The average year-to-year changes seen in Figure 8.12 reflect this, with the changes apparent around the ventricles and Sylvian fissure being stronger in LMCI and AD than CN and EMCI. We do however observe the highest year-to-year rate of change in LMCI, whereas one might expect this to be in AD. One potential reason for this is that the ADNI cohorts tend to include only mild cases of AD, as patients are often removed from the study once the severity

of their AD increases. As a result, the dataset only contains images from patients within their first few years of AD and therefore does not include examples covering the full range of atrophy levels, leading to such changes being underestimated by the model.

8.4.2 Effects of apolipoprotein E in LMCI and AD

Figure 8.13 shows a clear difference in the average year-to-year changes between $\epsilon4^-$ and $\epsilon4^+$ subjects, with an obviously increased rate of change in $\epsilon4^+$ subjects. The bootstrapping procedure gives us further confidence in the result. The method is shown to be stable, with the average results seen in Figure 8.14 appearing similar to those shown in Figure 8.13, and with a sufficiently small error range to indicate the difference between the two groups with a high degree of confidence. These significantly different regions correlate well with those areas associated with atrophy in AD - most obviously around the ventricles, as well as around the Sylvian fissure. This reflects the findings in [Nestor et al., 2008] of an increased rate of ventricular enlargement in $\epsilon4^+$ subjects. While the exact pathological mechanism for the effect of $\epsilon4$ in AD is unclear, it is known to effect Aβ aggregation and is therefore linked to the same downstream effects [Kim et al., 2009]. It, therefore, makes sense that $\epsilon4^+$ subjects will show an increased rate of change in other areas as well as the ventricles, an assertion supported by [Chalovich and Eisenberg, 2013], which reports that $\epsilon4$ carriers suffer an accelerated reduction in cortical thickness in both health and disease. The increased rate of change around the Sylvian fissure observed in $\epsilon4^+$ subjects is therefore not unexpected.

8.4.3 Comparison to direct image analysis

When modelling changes in pixel intensity directly in image space, we see an almost identical pattern of year-to-year changes across each disease state in Figure 8.15, with AD again showing less change than LMCI. However, this method fails to fully detect the difference in the year-to-year changes between $\epsilon4^-$ and $\epsilon4^+$ subjects. Average year-to-year changes shown in Figure 8.16 show little difference between the two groups, with only a slight indication of increased changes

around the temporal horns of the lateral ventricles in $\epsilon 4^+$ subjects. The regions of significant difference observed through bootstrapping appears noisy with little evidence of an increased rate of atrophy even around the ventricles. This demonstrates that while the simple analysis in image space is sufficient to detect the most significant differences between disease states, such as relative rates of atrophy, it fails to reliably detect subtler effects, such as that of APOE genotype which was readily identified through the via space.

This can be explained by considering that for each age, the image space model uses only a subset of the available images, weighted according to their ages relative to the target age. This means that each image only contributes to a small region of the model. However, the latent space model allows for a single image to be projected across all ages. The process of forming an atlas at a specific age then uses information from across the entire dataset. This means that the model at each age incorporates all examples of individual anatomical variation present in the dataset. This effectively averages out the natural and irrelevant variations in anatomy, reducing the noise and further exposing the effect of ageing.

Modelling in a latent space also has a number of advantages over image space beyond this increased sensitivity:

- Modelling in a latent space allows for specific effects to be visualised directly on an image as in Figure 8.10, rather than only predicting the average changes in intensity. This can be useful in establishing in precisely what way these changes manifest themselves, for example, whether these changes happen smoothly or suddenly.

- Using a latent space also allows for more complex relationships to be broken down. The exact anatomy of an individual's brain is a function of numerous genetic, disease and age-related factors, on top of their own unique basic pattern of cortical folding defined during the early stages of growth. A perfect model in latent space could account for all of these, ie, $z_{observed} = z_{basic} + z_{genes} + z_{disease} + z_{age}$.... With sufficient data, the effects of each of these factors could be isolated and measured through simple arithmetic. We have demonstrated how this allows longitudinal predictions to be made using cross-sectional data by projecting each image across the entire range of ages.

8.4.4 Applications

Being able to visualise the changes directly from images removes the need for an additional, and often imperfect, feature extraction step. For example, in [Nestor et al., 2008], the authors segment the ventricles to find their volume and use this feature to measure the effects of different APOE genotypes, thereby identifying the connection between the rate of ventricular atrophy and the presence of one or more $\epsilon 4$ alleles. The proposed method has detected the same effect and also suggests an increased rate of atrophy around the Sylvian fissure. To identify this in image space, a specific feature to measure this atrophy would need to be crafted and extracted before a potential relationship between this feature and APOE genotype can be established. Even assuming that such a feature could be consistently and accurately extracted, it would be a time-consuming process which is unlikely to happen unless there is prior knowledge to suggest such a relationship indeed exists.

However, the proposed method can only indicate expected changes in pixel intensity, and therefore provides no quantitative measure of these changes. We, therefore, see this method as a way to highlight areas which could be subject to further investigation. Once the overhead of GAN training and image encoding is carried out, it is a simple and fast process to visualise the changes in an image dataset associated with any disease state or clinical variable. This allows for fast, unsupervised, exploratory investigations into the relationships between any variable and changes seen in the imaging data. Once a potential effect has been identified, further feature based investigations could be carried out to confirm and quantify the effect.

8.5 Conclusion

In Part B of this chapter, we have demonstrated how the techniques proposed in Part A can be applied using a different GAN architecture to analyse and generate larger images. We showed that by performing analysis and producing models based upon the location of images in a latent space, associations can be learned between disease state and image characteristics. These associations were then projected onto synthetic images to visualise how the brain changes

in different AD related disease states. By incorporating age into the model, predictions of year-to-year changes under each disease state were made, allowing for the rate of progression to be compared between states, both at a population level, and when projected onto individual MR images.

Finally, we isolated the effect of a subject's APOE genotype to examine the difference in year-to-year changes between those with no $\epsilon 4$ allele and those with one or more, observing effects consistent with the literature. When compared to the results using a simple model in image space, we found a much greater sensitivity when modelling in latent space. We believe that performing latent space analysis in this way provides a useful tool for detecting associations between image features and disease state or clinical variables.

The drawback of modelling in a latent space is the need for an effective mapping between the latent and image spaces. On the evidence of the results shown in both parts of this chapter, GANs can provide an effective way of doing this. WGAN has been shown to learn a highly representative manifold for lower resolution images, allowing for a near perfect mapping between latent and image spaces. The PGGAN can be used at higher images sizes, but this comes at a cost of learning a less representative manifold, where mapping to and from this manifold does not always preserve high-frequency features such as cortical folding patterns. Despite this, we have shown that the latent space is expressive enough to be used to learn the relationships between clinical features and low-frequency imaging features. However, effects on higher frequency features such as local regions of increased atrophy are likely missed as they are not reliably encoded.

While we have shown that a given image can be projected across the age range, there is no guarantee that these images are accurate representations of how that particular brain would age, rather, we simply produce a prediction based upon the average changes observed over the rest of the data. This assumes that everyone ages at the same rate and that the physical changes associated with ageing manifest at the same ages. While this has proven to be adequate to get the results described here, improvements could be possible with better individual predictions. As discussed in Part A, this could be achieved by incorporating more individual

patient information into the model.

Future work will involve using different GAN architectures, or to improve the PGGAN architecture/training procedure, to allow for better mappings between latent and image spaces at high image sizes, with the ultimate aim of extending into 3D volumes.

Chapter 9

Future Work

Medical image computing, and in particular medical image synthesis, is a fast-moving field, with each step forwards opening up many new avenues for exploration. This chapter shares some ideas for such areas for investigation emanating from the work described so far. In addition, we share a couple of early attempts at producing 3D images using GANs, an important step towards the wider application of GANs within medical imaging. While neither method presented is currently suitable for applied usage, we hope that they may provide a significant starting point along two potential paths to 3D image generation.

9.1 Developments and extensions

This section aims to expand on the avenues for future work already discussed in each of the previous chapters by providing some alternative approaches and suggestions in greater detail.

9.1.1 Brain Lesion Segmentation through Image Synthesis and Outlier Detection

As mentioned in Section 4.6, there are several ways in which the proposed method can be improved, or extended to cover other pathologies. We demonstrated that comparisons with

pseudo-healthy images can be an effective way of detecting pathology, however, the task of realistic pseudo-healthy image synthesis remains a challenge. Our proposed kernel regression approach performs this task well, nevertheless, this comes at the cost of poor visual quality compared to other approaches. There are several ways in which GANs could be used to aid in the generation of pseudo-healthy images:

- Post-process the images generated by our proposed kernel regression approach to make them appear more realistic. This can be done in a similar way to [Shrivastava et al., 2016], where they improved the quality of their simulated images by using a network with an adversarial loss.

- Aggressively remove potentially pathological regions using a highly sensitive segmentation algorithm such as Lesion Prediction Algorithm (LST-LPA) with a low threshold. These missing regions can then be imputed through the method for image inpainting presented in [Yeh et al., 2016], using a GAN trained on healthy Fluid-attenuated Inversion Recovery (FLAIR) images. Such an approach will also remove the need for T_1-weighted images.

Either of these methods could lead to more accurate pseudo-healthy images and perhaps reduce the amount of post-processing required by eliminating more false positives and allowing for higher thresholds to be used. Another method for pseudo-healthy image synthesis has been proposed since the work in this thesis [Schlegl et al., 2017] and very recently explored further [Chen et al., 2018b, Chen and Konukoglu, 2018] by using generative models to learn a manifold of healthy images and constraining the reconstruction of a pathological image to lie on this manifold - thereby appearing healthy. Abnormalities can then be found through subtraction. While the generated images MR shown in [Chen et al., 2018b, Chen and Konukoglu, 2018] do not seem significantly higher quality than the ones generated by our proposed method (poor resemblance to the real images, though with no pathology), this is an interesting approach which warrants further investigation.

9.1.2 GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks

Using GANs for augmentation is still relatively new, and as such, there is a lot more to be explored. We have presented several techniques for GAN augmentation for segmentation tasks, whilst others have shown its uses in classification [Amitai and Goldberger, 2018, Moradi et al., 2018] tasks. In both of these applications, the first step should be to take the proposed methods and apply them to many different cases across anatomy and modalities to better understand how these approaches generalise.

9.1.3 GANsfer Learning: Combining labelled and unlabelled data for GAN based data augmentation

The introduction of unlabelled data into GAN augmentation is also something which can be explored further. We presented one method to do this, however, there are many others which can also be investigated.

One alternative could be to use a generative adversarial *transformation* network to effectively learn a set of transformations which could be used to deform a given image into a number of other realistic images. As discussed, the main problem with using deformation augmentation in medical imaging is that it is difficult to design a set of deformations which result in realistic images, indeed a deformation which is valid for one image may not be for another. It may be possible to learn a set of subject-specific deformations which can be used to transform a given image in a number of valid ways.

The network would be trained on unlabelled images, with the role of the generator to take as input a real image and output a realistically deformed version. The role of the discriminator would be, as always, to ensure the output of the generator has a realistic appearance. With sufficient regularisation to ensure the generator doesn't just learn an identity mapping, the generator could be then be used to provide a transformation to map a given labelled image onto

a point in the learned manifold of unlabelled images. This transformation can also be applied to the label channels, allowing for synthetic labelled data to be generated. Such an approach, if possible, would reduce the burden on the generator to generate real images from scratch. By only learning a transformation, and relying on the input image for the finer details, the resulting images should be of high quality. Similar approaches which aim to learn a series of realistic augmentations have already been proposed in [Lemley et al., 2017] and [Cubuk et al., 2018].

9.1.4 Modelling the Progression of Alzheimer's Disease in MRI Using Generative Adversarial Networks

Section 7.4 briefly describes some direct improvements which could be made to the process of generating subject-specific predictions, however, there are other avenues which could be explored by using the techniques described in Chapter 7. In particular, the proposed method has the potential to be a powerful tool in unsupervised data exploration. A proof of concept for this was provided in Chapter 8, with promising results. Large imaging datasets such as UKBiobank provide a valuable opportunity to measure correlations and associations between many clinical variables and imaging features. Exploring such datasets by looking for associations between a GAN derived latent space and clinical variables could yield some interesting discoveries. However, we identified some limitations in using the PGGAN algorithm to encode large images. If such image sizes are required, as is likely, more work must also be done to address some of these limitations.

9.2 Higher resolutions and 3D GANs

Common to all of the GAN based techniques presented in this thesis is the desire to go to higher image sizes and into 3D. One of the biggest drawbacks of current GAN formulations is their poor scaling to larger image sizes which has limited their potential in the applications discussed here. Both AD modelling and synthetic training data generation would benefit substantially

from being extended to 3D. While PGGAN has been shown to generate images up to 1024-by-1024, this is associated with 3-week training times on top-of-the-range hardware, and still contains a significantly smaller number of pixels than would be required to generate 3D 1mm isotropic brain images. The memory requirements required for such a high-resolution GAN are well beyond what is routinely available in modern hardware, and as such will require a different approach to simply increasing network size and compute power. In addition, current GAN formulations are constrained to simple, rectangular images due to their convolutional nature. This makes them unsuited to learning other types of data distributions common in medical images such as surface data.

Parallel to the work presented in this thesis, we have also investigated methods for extending GANs towards larger images sizes and into 3D. While none of these methods are mature enough to be used in any of the applications presented in this thesis, they will form the basis of future work.

We have decided to share the preliminary work for the most mature of these methods, in the hope that it may inspire future work in this direction. This approach involves generating 2-dimensional (2D), 3D or surface data at arbitrary resolutions by stitching together the output of a number of individually trained smaller GANs. The full details of this can be found in Appendix C.

Chapter 10

Conclusion

This thesis has described a number of contributions to the field of medical image synthesis. Chapter 4 presented a novel image synthesis technique which was particularly suited to the task of pseudo-healthy image synthesis. The proposed method was able to perform well at lesion segmentation, outperforming several popular methods.

Another use of synthetic data to aid in lesion segmentation, among other segmentation tasks, was investigated in Chapter 5. Here we demonstrated how using GAN derived synthetic training data to augment CNN training datasets can lead to significant improvements in performance, particularly in cases of limited data. We then built on this by showing how unlabelled data can be incorporated into the synthetic data generation process to further increase the amount of useful information which can be added to the training set. This was extended further in Chapter 6 where unlabelled data was introduced in a novel GAN training procedure. This improved the process of GAN augmentation by introducing synthetic labelled images with the characteristics of older and more pathological subjects than were present in the labelled dataset. This led to improved segmentation of images from elderly subjects and those suffering from AD, resulting in improved differentiation between mild and more severe cases of AD,

An alternative use of GANs, and the latent representations they learn, was presented in Chapters 7 and 8. Here, we demonstrated that certain image characteristics can be isolated and added or removed from images. Whilst we demonstrated this in the context of AD, the method

has the potential to be used in other applications, either as a method to generate subject specific synthetic images, or to identify the modes of variation between two groups of images. The technique, originally developed using the WGAN framework, was re-implemented using PG-GAN and applied to larger images. Further investigations were carried out, demonstrating the potential for the method to identify associations between clinical variables and image features. Finally, some avenues for future research were discussed in Chapter 9, including some early experimental results for the task of generating larger, 3D and surface image data.

Bibliography

[Abdulkadir et al., 2011] Abdulkadir, A., Mortamet, B., Vemuri, P., Jack, C. R., Krueger, G., and Klöppel, S. (2011). Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier. *NeuroImage*, 58(3):785–792.

[Adak et al., 2004] Adak, S., Illouz, K., Gorman, W., Tandon, R., Zimmerman, E. A., Guariglia, R., Moore, M. M., and Kaye, J. A. (2004). Predicting the rate of cognitive decline in aging and early Alzheimer disease among normal elderly individuals and patients with. *Neurol*, 63(1):108–114.

[Aharon et al., 2006] Aharon, M., Elad, M., and Bruckstein, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322.

[Aisen et al., 2010] Aisen, P. S., Petersen, R. C., Donohue, M. C., Gamst, A., Raman, R., Thomas, R. G., Walter, S., Trojanowski, J. Q., Shaw, L. M., Beckett, L. A., Jack, C. R., Jagust, W., Toga, A. W., Saykin, A. J., Morris, J. C., Green, R. C., and Weiner, M. W. (2010). Clinical core of the Alzheimer's disease neuroimaging initiative: Progress and plans. *Alzheimer's and Dementia*, 6(3):239–246.

[Alex et al., 2017] Alex, V., Safwan K. P., M., Chennamsetty, S. S., and Krishnamurthi, G. (2017). Generative adversarial networks for brain lesion detection. In *Medical Imaging 2017: Image Processing*, volume 10133, page 101330G. International Society for Optics and Photonics.

[Alexander et al., 2014] Alexander, D. C., Zikic, D., Zhang, J., Zhang, H., and Criminisi, A. (2014). Image quality transfer via random forest regression: applications in diffusion mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 225–232. Springer.

[Alexi et al., 2018] Alexi, B., Altuna, H., Babak, B. E., Wauters Carla, Geert, L., Jeroen, L. V., Dijk Van Marcory, Maschenka, B., Meyke, H., Nikolas, S., Oscar, G., Paul, D. V., Peter, B., Bult Peter, Manson Quirine, Vogels Rob, and Rob, D. L. V. (2018). Supporting data for "1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset".

[Alzheimer's Disease Neuroimaging Initiative, 2010] Alzheimer's Disease Neuroimaging Initiative (2010). Alzheimer's disease neuroimaging protocol grand opportunity (adni-go). `http://adni.loni.usc.edu/wp-content/themes/freshnews-dev-v2/documents/clinical/ADNI_Go_Protocol.pdf`.

[Amitai and Goldberger, 2018] Amitai, M. and Goldberger, J. (2018). Synthetic Data Augmentation Using GAN for Improved Liver Lesion Classification. *arXiv preprint arXiv:1801.02385*, pages 1–5.

[Ángel Muñoz-Ruiz et al., 2016] Ángel Muñoz-Ruiz, M., Hall, A., Mattila, J., Koikkalainen, J., Herukka, S.-K., Husso, M., Hänninen, T., Vanninen, R., Liu, Y., Hallikainen, M., Lötjönen, J., Remes, A. M., Alafuzoff, I., Soininen, H., and Hartikainen, P. (2016). Using the Disease State Fingerprint Tool for Differential Diagnosis of Frontotemporal Dementia and Alzheimer's Disease. *Dement Geriatr Cogn Disord Extra*, 66(2):313–329.

[Antoniou et al., 2017] Antoniou, A., Storkey, A., and Edwards, H. (2017). Data Augmentation Generative Adversarial Networks. *arXiv preprint arXiv:1711.04340*.

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv preprint arXiv:1701.07875*.

[Azami et al., 2016] Azami, M. E., Hammers, A., Jung, J., Costes, N., Bouet, R., and Lartizien, C. (2016). Detection of lesions underlying intractable epilepsy on T1-weighted MRI as an outlier detection problem. *PLoS ONE*, 11(9):e0161498.

[Babalola et al., 2008] Babalola, K. O., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., Smith, S., Cootes, T. F., Jenkinson, M., and Rueckert, D. (2008). Comparison and evaluation of segmentation techniques for subcortical structures in brain MRI. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5241 LNCS(PART 1):409–416.

[Balafar et al., 2010] Balafar, M. A., Ramli, A. R., Saripan, M. I., and Mashohor, S. (2010). Review of brain MRI image segmentation methods. *Artificial Intelligence Review*, 33(3):261–274.

[Barber et al., 1999] Barber, R., Scheltens, P., Gholkar, A., Ballard, C., McKeith, I., Ince, P., Perry, R., and O'Brien, J. (1999). White matter lesions on magnetic resonance imaging in dementia with Lewy bodies, Alzheimer's disease, vascular dementia, and normal aging. *Journal of Neurology Neurosurgery and Psychiatry*, 67(1):66–72.

[Barkhof et al., 2011] Barkhof, F., Fox, N. C., Bastos-Leite, A. J., and Scheltens, P. (2011). Normal Ageing. In *Neuroimaging in Dementia*, pages 43–57. Springer Berlin Heidelberg.

[Bateman et al., 2011] Bateman, R. J., Aisen, P. S., De Strooper, B., Fox, N. C., Lemere, C. A., Ringman, J. M., Salloway, S., Sperling, R. A., Windisch, M., and Xiong, C. (2011). Autosomal-dominant Alzheimer's disease: A review and proposal for the prevention of Alzheimer's disease. *Alzheimer's Research and Therapy*, 2(6):1.

[Belkin and Niyogi, 2003] Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396.

[Ben-Cohen et al., 2017] Ben-Cohen, A., Klang, E., Raskin, S. P., Amitai, M. M., and Greenspan, H. (2017). Virtual PET Images from CT Data Using Deep Convolutional Networks: Initial Results. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 49–57. Springer.

[Berthelot et al., 2017] Berthelot, D., Schumm, T., and Metz, L. (2017). BEGAN: Boundary Equilibrium Generative Adversarial Networks. *arXiv preprint arXiv:1703.10717*.

[Bi et al., 2017] Bi, L., Kim, J., Kumar, A., Feng, D., and Fulham, M. (2017). Synthesis of Positron Emission Tomography (PET) Images via Multi-channel Generative Adversarial Networks (GANs). In *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*, pages 43–51. Springer.

[Bland and Altman, 2010] Bland, J. M. and Altman, D. G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies*, 47(8):931–936.

[Bowles et al., 2018] Bowles, C., Gunn, R., Hammers, A., and Rueckert, D. (2018). Modelling the progression of Alzheimer's disease in MRI using generative adversarial networks. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105741K. International Society for Optics and Photonics.

[Bowles et al., 2017] Bowles, C., Qin, C., Guerrero, R., Gunn, R., Hammers, A., Dickie, D. A., Hernández, M. V., Wardlaw, J., and Rueckert, D. (2017). Brain lesion segmentation through image synthesis and outlier detection. *NeuroImage: Clinical*, 16:643–658.

[Boykov and Funka-Lea, 2006] Boykov, Y. and Funka-Lea, G. (2006). Graph Cuts and Efficient N-D Image Segmentation. *International Journal of Computer Vision*, 70(2):109–131.

[Buades et al., 2005] Buades, A., Coll, B., and Morel, J.-M. (2005). A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE.

[Burgos et al., 2014] Burgos, N., Cardoso, M. J., Thielemans, K., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Ahmed, R., Mahoney, C. J., Schott, J. M., Duncan, J. S., Atkinson, D., Arridge, S. R., Hutton, B. F., and Ourselin, S. (2014). Attenuation correction synthesis for hybrid PET-MR scanners: Application to brain studies. *IEEE Transactions on Medical Imaging*, 33(12):2332–2341.

[Burton et al., 2002] Burton, E., Karas, G., Paling, S., Barber, R., Williams, E., Ballard, C., McKeith, I., Scheltens, P., Barkhof, F., and O'Brien, J. (2002). Patterns of Cerebral Atrophy in Dementia with Lewy Bodies Using Voxel-Based Morphometry. *NeuroImage*, 17(2):618–630.

[Cajanus et al., 2018] Cajanus, A., Hall, A., Koikkalainen, J., Solje, E., Tolonen, A., Urhemaa, T., Liu, Y., Haanpää, R. M., Hartikainen, P., Helisalmi, S., Korhonen, V., Rueckert, D., Hasselbalch, S., Waldemar, G., Mecocci, P., Vanninen, R., van Gils, M., Soininen, H., Lötjönen, J., and Remes, A. M. (2018). Automatic MRI Quantifying Methods in Behavioral-Variant Frontotemporal Dementia Diagnosis. *Dementia and Geriatric Cognitive Disorders Extra*, 8(1):51–59.

[Caligiuri et al., 2015] Caligiuri, M. E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., and Cherubini, A. (2015). Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review. *Neuroinformatics*, 13(3):261–276.

[Cao et al., 2013] Cao, T., Jojic, V., Modla, S., Powell, D., Czymmek, K., and Niethammer, M. (2013). Robust multimodal dictionary learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 259–266. Springer.

[Cao et al., 2014] Cao, T., Zach, C., Modla, S., Powell, D., Czymmek, K., and Niethammer, M. (2014). Multi-modal registration for correlative microscopy using image analogies. *Medical Image Analysis*, 18(6):914–926.

[Cardoso et al., 2015] Cardoso, M. J., Sudre, C. H., Modat, M., and Ourselin, S. (2015). Template-based multimodal joint generative model of brain data. In *International conference on information processing in medical imaging*, pages 17–29. Springer.

[Caroli and Frisoni, 2010] Caroli, A. and Frisoni, G. B. (2010). The dynamics of Alzheimer's disease biomarkers in the Alzheimer's Disease Neuroimaging Initiative cohort. *Neurobiology of Aging*, 31(8):1263–1274.

[Casanova et al., 2011] Casanova, R., Whitlow, C. T., Wagner, B., Williamson, J., Shumaker, S. A., Maldjian, J. A., and Espeland, M. A. (2011). High Dimensional Classification of Structural MRI Alzheimer's Disease Data Based on Large Scale Regularization. *Frontiers in Neuroinformatics*, 5:22.

[Ceruti et al., 2012] Ceruti, C., Bassis, S., Rozza, A., Lombardi, G., Casiraghi, E., and Campadelli, P. (2012). DANCo: Dimensionality from angle and norm concentration. *arXiv preprint arXiv:1206.3881*.

[Chalovich and Eisenberg, 2013] Chalovich, J. M. and Eisenberg, E. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisms, and therapy. *Magn Reson Imaging*, 31(3):477–479.

[Chartsias et al., 2017] Chartsias, A., Joyce, T., Dharmakumar, R., and Tsaftaris, S. A. (2017). Adversarial image synthesis for unpaired multi-modal cardiac data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 3–13. Springer.

[Chartsias et al., 2018] Chartsias, A., Joyce, T., Giuffrida, M. V., and Tsaftaris, S. A. (2018). Multimodal MR Synthesis via Modality-Invariant Latent Representation. *IEEE Transactions on Medical Imaging*, 37(3):803–814.

[Chen et al., 2018a] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., and Rueckert, D. (2018a). Drinet for medical image segmentation. *IEEE transactions on medical imaging*, 37(11):2453–2462.

[Chen et al., 2015a] Chen, L., Tong, T., Ho, C. P., Patel, R., Cohen, D., Dawson, A. C., Halse, O., Geraghty, O., Rinne, P. E., White, C. J., et al. (2015a). 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.

[Chen et al., 2015b] Chen, M., Jog, A., Carass, A., and Prince, J. L. (2015b). Using image synthesis for multi-channel registration of different image modalities. In *Medical Imaging 2015: Image Processing*, volume 9413, page 94131Q. International Society for Optics and Photonics.

[Chen and Konukoglu, 2018] Chen, X. and Konukoglu, E. (2018). Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*.

[Chen et al., 2018b] Chen, X., Pawlowski, N., Rajchl, M., Glocker, B., and Konukoglu, E. (2018b). Deep Generative Models in the Real-World: An Open Challenge from Medical Imaging. *arXiv preprint arXiv:1806.05452*.

[Chen et al., 2018c] Chen, Y., Shi, F., Christodoulou, A. G., Zhou, Z., Xie, Y., and Li, D. (2018c). Efficient and Accurate MRI Super-Resolution using a Generative Adversarial Network and 3D Multi-Level Densely Connected Network. *arXiv preprint arXiv:1803.01417*.

[Chuquicusma et al., 2018] Chuquicusma, M. J., Hussein, S., Burt, J., and Bagci, U. (2018). How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 240–244. IEEE.

[Clark et al., 2007] Clark, C. M., Davatzikos, C., Borthakur, A., Newberg, A., Leight, S., Lee, V. M., and Trojanowski, J. Q. (2007). Biomarkers for early detection of Alzheimer pathology. *NeuroSignals*, 16(1):11–18.

[Costafreda et al., 2011] Costafreda, S. G., Dinov, I. D., Tu, Z., Shi, Y., Liu, C. Y., Kloszewska, I., Mecocci, P., Soininen, H., Tsolaki, M., Vellas, B., Wahlund, L. O., Spenger, C., Toga, A. W., Lovestone, S., and Simmons, A. (2011). Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. *NeuroImage*, 56(1):212–219.

[Cubuk et al., 2018] Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2018). Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

[Dai et al., 2017] Dai, W., Doyle, J., Liang, X., Zhang, H., Dong, N., Li, Y., and Xing, E. P. (2017). SCAN: Structure Correcting Adversarial Network for Organ Segmentation in Chest X-rays. *arXiv preprint arXiv:1703.08770*.

[Dawant et al., 2012] Dawant, B. M., Christensen, G. E., Fitzpatrick, J. M., and Rueckert, D., editors (2012). *Biomedical Image Registration*, volume 7359 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

[Debette and Markus, 2010] Debette, S. and Markus, H. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *Bmj*, 341:c3666.

[Del C. Valdes Hernandez et al., 2013] Del C. Valdes Hernandez, M., Piper, R. J., Wang, X., Deary, I. J., and Wardlaw, J. M. (2013). Towards the automatic computational assessment of enlarged perivascular spaces on brain magnetic resonance images: A systematic review. *Journal of Magnetic Resonance Imaging*, 38(4):774–785.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (Methodological)*, 39(1):1–38.

[Denton et al., 2015] Denton, E., Chintala, S., Szlam, A., and Fergus, R. (2015). Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Advances in neural information processing systems*, pages 1486–1494.

[Dice, 1945] Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

[Donoho, 2006] Donoho, D. L. (2006). Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306.

[Doody et al., 2001] Doody, R. S., Massman, P., and Dunn, K. (2001). A method for estimating pregression rates in alzheimer disease. *Arch. Neurol.*, 58(3):449–454.

[Dosovitskiy et al., 2015] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., and Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766.

[Dukart et al., 2011] Dukart, J., Mueller, K., Horstmann, A., Barthel, H., Möller, H. E., Villringer, A., Sabri, O., and Schroeter, M. L. (2011). Combined evaluation of FDG-PET and MRI improves detection and differentiation of dementia. *PLoS ONE*, 6(3):e18111.

[Dwarikanath et al., 2018] Dwarikanath, M., Antony, B., Sedai, S., and Garnavi, R. (2018). Deformable medical image registration using generative adversarial networks. In *IEEE International Symposium on Biomedical Imaging*, pages 1449–1453.

[Farrell et al., 2009] Farrell, C., Chappell, F., Armitage, P. A., and Keston, P. (2009). Development and initial testing of normal reference MR images for the brain at ages. *European Radiology*, 19(1):177–183.

[Farrer, 1997] Farrer, L. A. (1997). Effects of Age, Sex, and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease. *Jama*, 278(16):1349.

[Fischl et al., 2004] Fischl, B., Salat, D. H., Van Der Kouwe, A. J., Makris, N., Ségonne, F., Quinn, B. T., and Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *Neuroimage*, 23:S69–S84.

[Foster et al., 2007] Foster, N. L., Heidebrink, J. L., Clark, C. M., Jagust, W. J., Arnold, S. E., Barbas, N. R., DeCarli, C. S., Scott Turner, R., Koeppe, R. A., Higdon, R., and Minoshima, S. (2007). FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease. *Brain*, 130(10):2616–2635.

[Freeborough and Fox, 1997] Freeborough, P. A. and Fox, N. C. (1997). The boundary shift integral: An accurate and robust measure of cerebral volume changes from registered repeat MRI. *IEEE Transactions on Medical Imaging*, 16(5):623–629.

[Freund and Schapire, 1996] Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *ICML '96: Proceedings of the 13th International Conference on Machine Learning*, 96:148–156.

[Fu et al., 1998] Fu, L., Matthews, P. M., De Stefano, N., Worsley, K. J., Narayanan, S., Francis, G. S., Antel, J. P., Wolfson, C., and Arnold, D. L. (1998). Imaging axonal damage of normal-appearing white matter in multiple sclerosis. *Brain*, 121(1):103–113.

[Fukunaga and Hostetler, 1975] Fukunaga, K. and Hostetler, L. (1975). The Estimation of the Gradient of a Density Function. *IEEE Transactions on Information Theory*, 21(1):32–40.

[Galton et al., 2001] Galton, C. J., Patterson, K., Graham, K., Lambon-Ralph, M. A., Williams, G., Antoun, N., Sahakian, B. J., and Hodges, J. R. (2001). Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. *Neurology*, 57(2):216–225.

[Garcia-Lorenzo et al., 2013] Garcia-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L., and Collins, D. L. (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical image analysis*, 17(1):1–18.

[García-Lorenzo et al., 2009] García-Lorenzo, D., Lecoeur, J., Arnold, D. L., Collins, D. L., and Barillot, C. (2009). Multiple sclerosis lesion segmentation using an automatic multimodal graph cuts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 584–591. Springer.

[García-Lorenzo et al., 2008] García-Lorenzo, D., Prima, S., Collins, D. L., Arnold, D., Morrissey, S. P., and Barillot, C. (2008). Combining robust expectation maximization and mean shift algorithms for multiple sclerosis brain segmentation. In *MICCAI workshop on Medical Image Analysis on Multiple Sclerosis (validation and methodological issues)(MIAMS'2008)*, pages 82–91.

[Garcia-Lorenzo et al., 2008] Garcia-Lorenzo, D., Prima, S., Morrissey, S. P., and Barillot, C. (2008). A robust expectation-maximization algorithm for multiple sclerosis lesion segmentation. In *MICCAI Workshop: 3D Segmentation in the Clinic: A Grand Challenge II, MS Lesion Segmentation*, page 277.

[Genin et al., 2011] Genin, E., Hannequin, D., Wallon, D., Sleegers, K., Hiltunen, M., Combarros, O., Bullido, M. J., Engelborghs, S., De Deyn, P., Berr, C., Pasquier, F., Dubois, B., Tognoni, G., Fiévet, N., Brouwers, N., Bettens, K., Arosio, B., Coto, E., Del Zompo, M., Mateo, I., Epelbaum, J., Frank-Garcia, A., Helisalmi, S., Porcellini, E., Pilotto, A., Forti,

P., Ferri, R., Scarpini, E., Siciliano, G., Solfrizzi, V., Sorbi, S., Spalletta, G., Valdivieso, F., Vepsäläinen, S., Alvarez, V., Bosco, P., Mancuso, M., Panza, F., Nacmias, B., Boss, P., Hanon, O., Piccardi, P., Annoni, G., Seripa, D., Galimberti, D., Licastro, F., Soininen, H., Dartigues, J. F., Kamboh, M. I., Van Broeckhoven, C., Lambert, J. C., Amouyel, P., and Campion, D. (2011). APOE and Alzheimer disease: A major gene with semi-dominant inheritance. *Molecular Psychiatry*, 16(9):903–907.

[Ghafoorian et al., 2017] Ghafoorian, M., Karssemeijer, N., Heskes, T., Bergkamp, M., Wissink, J., Obels, J., Keizer, K., de Leeuw, F. E., van Ginneken, B., Marchiori, E., and Platel, B. (2017). Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. *NeuroImage: Clinical*, 14:391–399.

[Global Action Against Dementia, 2013] Global Action Against Dementia (2013). G8 Dementia Summit Declaration. `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/265869/2901668_G8_DementiaSummitDeclaration_acc.pdf`.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[Gray et al., 2013] Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., and Rueckert, D. (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage*, 65:167–175.

[Grewenig et al., 2011] Grewenig, S., Zimmer, S., and Weickert, J. (2011). Rotationally invariant similarity measures for nonlocal image denoising. *Journal of Visual Communication and Image Representation*, 22(2):117–130.

[Griffanti et al., 2016] Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U. G., Kuker, W., Battaglini, M., Rothwell, P. M., and Jenkinson, M. (2016). BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, 141:191–205.

[Guerreiro et al., 2013] Guerreiro, R., Wojtas, A., Bras, J., Carrasquillo, M., Rogaeva, E., Majounie, E., Cruchaga, C., Sassi, C., Kauwe, J. S., Younkin, S., Hazrati, L., Collinge, J., Pocock, J., Lashley, T., Williams, J., Lambert, J.-C., Amouyel, P., Goate, A., Rademakers, R., Morgan, K., Powell, J., St. George-Hyslop, P., Singleton, A., and Hardy, J. (2013). TREM2 Variants in Alzheimer's Disease. *New England Journal of Medicine*, 368(2):117–127.

[Guerrero et al., 2018] Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M. C., Dickie, D. A., Wardlaw, J., and Rueckert, D. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical*, 17:918–934.

[Gulrajani et al., 2017] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved Training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*.

[Hajnal et al., 1992] Hajnal, J. V., Bryant, D. J., Kasuboski, L., Pattany, P. M., Coene, B. D., Lewis, P. D., Pennock, J. M., Oatridge, A., Young, I. R., and Bydder, G. M. (1992). Use of fluid attenuated inversion recovery (FLAIR) pulse sequences in MRI of the brain. *Journal of Computer Assisted Tomography*, 16(6):841–844.

[Haller et al., 2016] Haller, S., Falkovskiy, P., Meuli, R., Thiran, J. P., Krueger, G., Lovblad, K. O., Kober, T., Roche, A., and Marechal, B. (2016). Basic MR sequence parameters systematically bias automated brain volume estimation. *Neuroradiology*, 58(11):1153–1160.

[Hampel et al., 2008] Hampel, H., Bürger, K., Teipel, S. J., Bokde, A. L., Zetterberg, H., and Blennow, K. (2008). Core candidate neurochemical and imaging biomarkers of alzheimers disease. *Alzheimer's & Dementia*, 4(1):38–48.

[Heckemann et al., 2010] Heckemann, R. A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J. V., and Hammers, A. (2010). Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. *NeuroImage*, 51(1):221–227.

[Heckemann et al., 2015] Heckemann, R. A., Ledig, C., Gray, K. R., Aljabar, P., Rueckert, D., Hajnal, J. V., and Hammers, A. (2015). Brain extraction using label propagation and group agreement: Pincram. *PLoS ONE*, 10(7):e0129211.

[Herholz, 1995] Herholz, K. (1995). FDG PET and differential diagnosis of dementia. *Alzheimer Disease and Associated Disorders*, 9(1):6–16.

[Hertzmann et al., 2001] Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., and Salesin, D. H. (2001). Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340. ACM.

[Hoffman et al., 2000] Hoffman, J. M., Welsh-Bohmer, K. A., Hanson, M., Crain, B., Hulette, C. M., Earl, N., and Coleman, R. E. (2000). FDG PET imaging in patients with pathologically verified dementia. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 41(11):1920–1928.

[Huang and Wang, 2013] Huang, D. A. and Wang, Y. C. F. (2013). Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2496–2503. IEEE.

[Huo et al., 2018] Huo, Y., Xu, Z., Bao, S., Bermudez, C., Plassard, A. J., Yao, Y., Liu, J., Assad, A., Abramson, R. G., and Landman, B. A. (2018). Splenomegaly segmentation using global convolutional kernels and conditional generative adversarial networks. In *Medical Imaging 2018: Image Processing*, volume 10574, page 8. International Society for Optics and Photonics.

[Huppertz et al., 2011] Huppertz, H. J., Wagner, J., Weber, B., House, P., and Urbach, H. (2011). Automated quantitative FLAIR analysis in hippocampal sclerosis. *Epilepsy Research*, 97(1-2):146–156.

[Iglesias et al., 2013] Iglesias, J. E., Konukoglu, E., Zikic, D., Glocker, B., Van Leemput, K., and Fischl, B. (2013). Is synthesizing mri contrast useful for inter-modality analysis? In

International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 631–638. Springer.

[Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

[Izadi et al., 2018] Izadi, S., Mirikharaji, Z., Kawahara, J., and Hamarneh, G. (2018). Generative Adversarial Networks To Segment Skin Lesions. In *International Symposium on Biomedical Imaging (ISBI)*, pages 881–884. IEEE.

[Jack et al., 1999] Jack, C. R., Petersen, R. C., Xu, Y. C., O'Brien, P. C., Smith, G. E., Ivnik, R. J., Boeve, B. F., Waring, S. C., Tangalos, E. G., and Kokmen, E. (1999). Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology*, 52(7):1397–1397.

[Jack et al., 2010] Jack, R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., and Trojanowski, J. Q. (2010). Personal View Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurology*, 9(1):119–128.

[Jenkinson et al., 2012] Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *NeuroImage*, 62(2):782–790.

[Jin et al., 2018] Jin, C.-B., Jung, W., Joo, S., Park, E., Saem, A. Y., Han, I. H., Lee, J. I., and Cui, X. (2018). Deep CT to MR Synthesis using Paired and Unpaired Data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 14–23. Springer.

[Jog et al., 2015] Jog, A., Carass, A., Pham, D. L., and Prince, J. L. (2015). Tree-encoded conditional random fields for image synthesis. In *International Conference on Information Processing in Medical Imaging*, pages 733–745. Springer.

[Jog et al., 2017] Jog, A., Carass, A., Roy, S., Pham, D. L., and Prince, J. L. (2017). Random forest regression for magnetic resonance image synthesis. *Medical Image Analysis*, 35:475–488.

[Jog et al., 2013] Jog, A., Roy, S., Carass, A., and Prince, J. L. (2013). Magnetic resonance image synthesis through patch regression. *IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, 2013:350–353.

[Joyce et al., 2018] Joyce, T., Chartsias, A., and Tsaftaris, S. (2018). Deep multi-class segmentation without ground-truth labels. In *Medical Imaging with Deep Learning*.

[Kamboh, 1995] Kamboh, M. I. (1995). Apolipoprotein E polymorphism and susceptibility to Alzheimer's disease. *Hum.Biol.*, 67(0018-7143):195–215.

[Kamnitsas et al., 2017a] Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al. (2017a). Ensembles of multiple models and architectures for robust brain tumour segmentation. In *International MICCAI Brainlesion Workshop*, pages 450–462. Springer.

[Kamnitsas et al., 2017b] Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. (2017b). Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. *Medical Image Analysis*, 36:61–78.

[Kang et al., 2012] Kang, X., Herron, T. J., Cate, A. D., Yund, E. W., and Woods, D. L. (2012). Hemispherically-Unified Surface Maps of Human Cerebral Cortex: Reliability and Hemispheric Asymmetries. *PLoS ONE*, 7(9):e45582.

[Kannan et al., 2009] Kannan, S., Balakrishnan, B., Muzik, O., Romero, R., and Chugani, D. (2009). Positron emission tomography imaging of neuroinflammation. *J Child Neurol*, 24(9):1190–1199.

[Karpate et al., 2015] Karpate, Y., Commowick, O., and Barillot, C. (2015). Probabilistic One Class Learning for Automatic Detection of Multiple Sclerosis Lesions. *IEEE International Symposium on Biomedical Imaging*, pages 486–489.

[Karras et al., 2017] Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196*.

[Kim et al., 2009] Kim, J., Basak, J. M., and Holtzman, D. M. (2009). The Role of Apolipoprotein E in Alzheimer's Disease. *Neuron*, 63(3):287–303.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Kohl et al., 2017] Kohl, S., Bonekamp, D., Schlemmer, H.-P., Yaqubi, K., Hohenfellner, M., Hadaschik, B., Radtke, J.-P., and Maier-Hein, K. (2017). Adversarial Networks for the Detection of Aggressive Prostate Cancer. *arXiv preprint arXiv:1702.08014*.

[Koikkalainen et al., 2016] Koikkalainen, J., Rhodius-meester, H., Tolonen, A., Barkhof, F., Tijms, B., Lemstra, W., Tong, T., Guerrero, R., Schuh, A., Ledig, C., Rueckert, D., Soininen, H., Remes, A. M., Waldemar, G., Hasselbalch, S., Mecocci, P., Flier, W. V. D., and Lötjönen, J. (2016). NeuroImage : Clinical Differential diagnosis of neurodegenerative diseases using structural MRI data. *NeuroImage: Clinical*, 11:435–449.

[Konukoglu et al., 2013] Konukoglu, E., van der Kouwe, A., Sabuncu, M. R., and Fischl, B. (2013). Example-based restoration of high-resolution magnetic resonance image acquisitions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 131–138. Springer.

[Krähenbühl and Koltun, 2011] Krähenbühl, P. and Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117.

[Krivov et al., 2017] Krivov, E., Pisov, M., and Belyaev, M. (2017). Mri augmentation via elastic registration for brain lesions segmentation. In *International MICCAI Brainlesion Workshop*, pages 369–380. Springer.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[Kroon and Slump, 2009] Kroon, D. J. and Slump, C. H. (2009). MRI modalitiy transformation in demon registration. In *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, pages 963–966. IEEE.

[Kuijf et al., 2012] Kuijf, H. J., de Bresser, J., Geerlings, M. I., Conijn, M. M., Viergever, M. A., Biessels, G. J., and Vincken, K. L. (2012). Efficient detection of cerebral microbleeds on 7.0T MR images using the radial symmetry transform. *NeuroImage*, 59(3):2266–2273.

[Kurz, 2001] Kurz, A. (2001). What is vascular dementia? *International journal of clinical practice. Supplement*, (120):5–8.

[Ledig et al., 2015] Ledig, C., Heckemann, R. A., Hammers, A., Lopez, J. C., Newcombe, V. F., Makropoulos, A., Lötjönen, J., Menon, D. K., and Rueckert, D. (2015). Robust whole-brain segmentation: Application to traumatic brain injury. *Medical Image Analysis*, 21(1):40–58.

[Ledig et al., 2018a] Ledig, C., Schuh, A., Guerrero, R., Heckemann, R. A., and Rueckert, D. (2018a). Structural brain imaging in alzheimers disease and mild cognitive impairment: biomarker analysis and shared morphometry database. *Scientific reports*, 8(1):11258.

[Ledig et al., 2018b] Ledig, C., Schuh, A., Guerrero, R., Heckemann, R. A., and Rueckert, D. (2018b). Structural brain imaging in alzheimers disease and mild cognitive impairment: biomarker analysis and shared morphometry database. *Scientific reports*, 8(1):11258.

[Ledig et al., 2016] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2016). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *arXiv preprint arXiv:1609.04802*.

[Lemley et al., 2017] Lemley, J., Bazrafkan, S., and Corcoran, P. (2017). Smart augmentation learning an optimal data augmentation strategy. *IEEE Access*, 5:5858–5869.

[Li et al., 2017] Li, Z., Wang, Y., and Yu, J. (2017). Brain tumor segmentation using an adversarial network. In *International MICCAI Brainlesion Workshop*, pages 123–132. Springer.

[Liu et al., 2013] Liu, Y., Mattila, J., Ruiz, M. Á. M., Paajanen, T., Koikkalainen, J., van Gils, M., Herukka, S. K., Waldemar, G., Lötjönen, J., and Soininen, H. (2013). Predicting

AD Conversion: Comparison between Prodromal AD Guidelines and Computer Assisted PredictAD Tool. *PLoS ONE*, 8(2):e55246.

[Lladó et al., 2012] Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J. C., Quiles, A., Valls, L., Ramió-Torrentà, L., and Rovira, À. (2012). Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches. *Information Sciences*, 186(1):164–185.

[Lombardi et al., 2011] Lombardi, G., Rozza, A., Ceruti, C., Casiraghi, E., and Campadelli, P. (2011). Minimum neighbor distance estimators of intrinsic dimension. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 374–389. Springer.

[Lucic et al., 2017] Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. (2017). Are GANs Created Equal? A Large-Scale Study. *arXiv preprint arXiv:1711.10337*.

[Maaten and Hinton, 2008] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

[Madani et al., 2018] Madani, A., Moradi, M., Karargyris, A., and Syeda-Mahmood, T. (2018). Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1038–1042. IEEE.

[Maier et al., 2018] Maier, A., Schebesch, F., Syben, C., Würfl, T., Steidl, S., Choi, J.-H., and Fahrig, R. (2018). Precision learning: towards use of known operators in neural networks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 183–188. IEEE.

[Mao et al., 2017] Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802.

[Mariani et al., 2018] Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., and Malossi, C. (2018). BAGAN: Data Augmentation with Balancing GAN. *arXiv preprint arXiv:1803.09655*.

[Mattila et al., 2011] Mattila, J., Koikkalainen, J., Virkki, A., Simonsen, A., Van Gils, M., Waldemar, G., Soininen, H., and Lötjönen, J. (2011). A disease state fingerprint for evaluation of alzheimer's disease. *Journal of Alzheimer's Disease*, 27(1):163–176.

[Mattila et al., 2012] Mattila, J., Koikkalainen, J., Virkki, A., Van Gils, M., and Lötjönen, J. (2012). Design and application of a generic clinical decision support system for multiscale data. *IEEE Transactions on Biomedical Engineering*, 59(1):234–240.

[Mazziotta et al., 1992] Mazziotta, J. C., Frackowiak, R. S., and Phelps, M. E. (1992). The use of positron emission tomography in the clinical assessment of dementia. *Seminars in Nuclear Medicine*, 22(4):233–246.

[McAuliffe et al., 2001] McAuliffe, M. J., Lalonde, F. M., McGarry, D., Gandler, W., Csaky, K., and Trus, B. L. (2001). Medical image processing, analysis and visualization in clinical research. In *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001*, pages 381–386. IEEE.

[McGraw, K., Wong, 1996] McGraw, K., Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.

[McKhann et al., 2011] McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., Mohs, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillo, M. C., Thies, B., Weintraub, S., and Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia*, 7(3):263–269.

[McLachlan et al., 2019] McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6:355–378.

[Miller et al., 1993] Miller, M. I., Christensen, G. E., Amit, Y., and Grenander, U. (1993). Mathematical textbook of deformable neuroanatomies. *Proceedings of the National Academy of Sciences*, 90(24):11944–11948.

[Milletari et al., 2016] Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE.

[Misra et al., 2009] Misra, C., Fan, Y., and Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *NeuroImage*, 44(4):1415–1422.

[Moeskops et al., 2017] Moeskops, P., Veta, M., Lafarge, M. W., Eppenhof, K. A., and Pluim, J. P. (2017). Adversarial training and dilated convolutions for brain MRI segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 56–64. Springer.

[Mok and Chung, 2018] Mok, T. C. W. and Chung, A. C. S. (2018). Learning Data Augmentation for Brain Tumor Segmentation with Coarse-to-Fine Generative Adversarial Networks. *arXiv preprint arXiv:1805.11291*.

[Moradi et al., 2018] Moradi, M., Madani, A., Karargyris, A., and Syeda-Mahmood, T. F. (2018). Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In *Medical Imaging 2018: Image Processing*, page 57. International Society for Optics and Photonics.

[Mueller et al., 2005] Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877.

[Nestor et al., 2008] Nestor, S. M., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., Wells, J. L., Fogarty, J., and Bartha, R. (2008). Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain*, 131(9):2443–2454.

[Nie et al., 2016] Nie, D., Trullo, R., Petitjean, C., Ruan, S., and Shen, D. (2016). Medical Image Synthesis with Context-Aware Generative Adversarial Networks. *arXiv preprint arXiv:1612.05362*.

[Odena et al., 2016] Odena, A., Olah, C., and Shlens, J. (2016). Conditional Image Synthesis With Auxiliary Classifier GANs. *arXiv preprint arXiv:1610.09585*.

[Polman et al., 2011] Polman, C. H., Reingold, S. C., Banwell, B., Clanet, M., Cohen, J. A., Filippi, M., Fujihara, K., Havrdova, E., Hutchinson, M., Kappos, L., Lublin, F. D., Montalban, X., O'Connor, P., Sandberg-Wollheim, M., Thompson, A. J., Waubant, E., Weinshenker, B., and Wolinsky, J. S. (2011). Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria. *Annals of Neurology*, 69(2):292–302.

[Potter et al., 2015] Potter, G. M., Chappell, F. M., Morris, Z., and Wardlaw, J. M. (2015). Cerebral perivascular spaces visible on magnetic resonance imaging: Development of a qualitative rating scale and its observer reliability. *Cerebrovascular Diseases*, 39(3-4):224–231.

[Prince, 2015] Prince, M. J. (2015). *World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease International.

[Quan et al., 2017] Quan, T. M., Nguyen-Duc, T., and Jeong, W.-K. (2017). Compressed Sensing MRI Reconstruction using a Generative Adversarial Network with a Cyclic Loss. *IEEE transactions on medical imaging*, 37(6):1488–1497.

[Rabin et al., 2012] Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2012). Wasserstein barycenter and its application to texture mixing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6667 LNCS, pages 435–446. Springer.

[Rabinovici et al., 2008] Rabinovici, G. D., Seeley, W. W., Kim, E. J., Gorno-Tempini, M. L., Rascovsky, K., Pagliaro, T. A., Allison, S. C., Halabi, C., Kramer, J. H., Johnson, J. K., Weiner, M. W., Forman, M. S., Trojanowski, J. Q., Dearmond, S. J., Miller, B. L., and Rosen, H. J. (2008). Distinct MRI atrophy patterns in autopsy-proven Alzheimer's disease and frontotemporal lobar degeneration. *American Journal of Alzheimer's Disease and other Dementias*, 22(6):474–488.

[Radford et al., 2015] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1511.06434*.

[Rezaei et al., 2017] Rezaei, M., Harmuth, K., Gierke, W., Kellermeier, T., Fischer, M., Yang, H., and Meinel, C. (2017). A conditional adversarial network for semantic segmentation of brain tumor. In *International MICCAI Brainlesion Workshop*, pages 241–252. Springer.

[Richter et al., 2016] Richter, S. R., Vineet, V., Roth, S., and Koltun, V. (2016). Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer.

[Román et al., 1993] Román, G. C., Tatemichi, T. K., Erkinjuntti, T., Cummings, J., Masdeu, J., Garcia, J., Amaducci, L., Orgogozo, J.-M., Brun, A., Hofman, A., et al. (1993). Vascular dementia: diagnostic criteria for research studies: report of the NINDS-AIREN International Workshop. *Neurology*, 43(2):250–250.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

[Ross et al., 2018] Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., Kenngott, H., Speidel, S., Kopp-Schneider, A., Maier-Hein, K., and Maier-Hein, L. (2018). Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International Journal of Computer Assisted Radiology and Surgery*, 13(6):925–933.

[Roy et al., 2017] Roy, S., Butman, J. A., and Pham, D. L. (2017). Synthesizing ct from ultrashort echo-time mr images via convolutional neural networks. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 24–32. Springer.

[Roy et al., 2014a] Roy, S., Carass, A., Jog, A., Prince, J. L., and Lee, J. (2014a). MR to CT registration of brains using image synthesis. *Proceedings of SPIE–the International Society for Optical Engineering*, 9034:903419.

[Roy et al., 2014b] Roy, S., Carass, A., Jog, A., Prince, J. L., and Lee, J. (2014b). Mr to ct registration of brains using image synthesis. In *Medical Imaging 2014: Image Processing*, volume 9034, page 903419. International Society for Optics and Photonics.

[Roy et al., 2011] Roy, S., Carass, A., and Prince, J. (2011). A compressed sensing approach for MR tissue contrast synthesis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6801 LNCS:371–383.

[Roy et al., 2013] Roy, S., Carass, A., and Prince, J. L. (2013). Magnetic resonance image example-based contrast synthesis. *IEEE Transactions on Medical Imaging*, 32(12):2348–2363.

[Roy et al., 2010] Roy, S., Carass, A., Shiee, N., Pham, D. L., and Prince, J. L. (2010). Mr contrast synthesis for lesion segmentation. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 932–935. IEEE.

[Rueckert et al., 1999] Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L., Leach, M. O., and Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: application to breast MR images. *IEEE transactions on medical imaging*, 18(8):712–721.

[Rueda et al., 2013] Rueda, A., Malpica, N., and Romero, E. (2013). Single-image super-resolution of brain MR images using overcomplete dictionaries. *Medical Image Analysis*, 17(1):113–132.

[Salehinejad et al., 2017] Salehinejad, H., Valaee, S., Dowdell, T., Colak, E., and Barfett, J. (2017). Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks. *arXiv preprint arXiv:1712.01636*.

[Sánchez and Vilaplana, 2018] Sánchez, I. and Vilaplana, V. (2018). Brain mri super-resolution using 3d generative adversarial networks. *arXiv preprint arXiv:1812.11440*.

[Schlegl et al., 2017] Schlegl, T., Seeböck, P., Waldstein, S., and Langs, G. (2017). Unsupervised Anomaly Detection with Marker Discovery. *Information Processing in Medical Imaging (IPMI)*, pages 146–157.

[Schmidt et al., 2012] Schmidt, P., Arsic, M., Buck, D., Forschler, A., Berthele, A., Gaser, C., Hoshi, M., Ilg, R., Schmid, V., Zimmer, C., Hemmer, B., and Muhlau, M. (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *NeuroImage*, 59(4):3774–83.

[Schuff et al., 2009] Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L. M., Trojanowski, J. Q., Thompson, P. M., Jack, C. R., and Weiner, M. W. (2009). MRI of hippocampal volume loss in early Alzheimers disease in relation to ApoE genotype and biomarkers. *Brain*, 132(4):1067–1077.

[Seab et al., 1988] Seab, J. P., Jagust, W. J., Wong, S. T., Roos, M. S., Reed, B. R., and Budinger, T. F. (1988). Quantitative NMR measurements of hippocampal atrophy in Alzheimer's disease. *Magnetic Resonance in Medicine*, 8(2):200–208.

[Sevetlidis et al., 2016] Sevetlidis, V., Giuffrida, M. V., and Tsaftaris, S. A. (2016). Whole image synthesis using a deep encoder-decoder network. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 127–137. Springer.

[Shankaranarayana et al., 2017] Shankaranarayana, S. M., Ram, K., Mitra, K., and Sivaprakasam, M. (2017). Joint optic disc and cup segmentation using fully convolutional and adversarial networks. In *Fetal, Infant and Ophthalmic Medical Image Analysis*, pages 168–176. Springer.

[Shaw et al., 2007] Shaw, L. M., Korecka, M., Clark, C. M., Lee, V. M., and Trojanowski, J. Q. (2007). Biomarkers of neurodegeneration for diagnosis and monitoring therapeutics. *Nature Reviews Drug Discovery*, 6(4):295–303.

[Shaw et al., 2009] Shaw, L. M., Vanderstichele, H., Knapik-Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., Blennow, K., Soares, H., Simon, A., Lewczuk, P., Dean, R., Siemers, E., Potter, W., Lee, V. M., and Trojanowski, J. Q. (2009). Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects. *Annals of Neurology*, 65(4):403–413.

[Shiee et al., 2010] Shiee, N., Bazin, P. L., Ozturk, A., Reich, D. S., Calabresi, P. A., and Pham, D. L. (2010). A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, 49(2):1524–1535.

[Shin et al., 2016] Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298.

[Shock et al., 1984] Shock, N. W., Greulich, R. C., Andres, R., Aenberg, D., Costa, P. T. J., Lakatta, E. G., and Tobin, J. D. (1984). Normal Human Aging: The Baltimore Longitudinal Study of Aging. page 661. `http://eric.ed.gov/?id=ED292030`.

[Shrivastava et al., 2016] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2016). Learning from Simulated and Unsupervised Images through Adversarial Training. *arXiv preprint arXiv:1612.07828*.

[Silverman et al., 2001] Silverman, D. H. S., Small, G. W., Chang, C. Y., Lu, C. S., de Aburto, M. A. K., Chen, W., Czernin, J., Rapoport, S. I., Pietrini, P., Alexander, G. E., Schapiro, M. B., Jagust, W. J., Hoffman, J. M., Welsh-Bohmer, K. A., Alavi, A., Clark, C. M., Salmon, E., de Leon, M. J., Mielke, R., Cummings, J. L., Kowell, A. P., Gambhir, S. S., Hoh, C. K., and Phelps, M. E. (2001). Positron Emission Tomography in Evaluation of Dementia. *Jama*, 286(17):2120.

[Son et al., 2017] Son, J., Park, S. J., and Jung, K.-H. (2017). Retinal Vessel Segmentation in Fundoscopic Images with Generative Adversarial Networks. *arXiv preprint arXiv:1706.09318*.

[Souplet et al., 2008] Souplet, J., Lebrun, C., Ayache, N., and Malandain, G. (2008). An automatic segmentation of multiple sclerosis lesions. `http://www.ia.unc.edu/MSseg/papers/IJ_613.pdf`.

[Styner et al., 2008] Styner, M., Lee, J., Chin, B., and Chin, M. (2008). 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. *Midas*, 2008:1–6.

[Suk et al., 2015] Suk, H. I., Lee, S. W., and Shen, D. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Structure and Function*, 220(2):841–859.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

[Tamura and Araki, 2015] Tamura, Y. and Araki, A. (2015). Diabetes mellitus and white matter hyperintensity. *Geriatrics and Gerontology International*, 15:34–42.

[Thompson et al., 2001] Thompson, P. M., Mega, M. S., Woods, R. P., Zoumalan, C. I., Lindshield, C. J., Blanton, R. E., Moussai, J., Holmes, C. J., Cummings, J. L., and Toga, A. W. (2001). Cortical change in Alzheimer's disease detected with a disease-specific population-based brain atlas. *Cerebral Cortex*, 11(1):1–16.

[Tolstikhin et al., 2017] Tolstikhin, I., Gelly, S., Bousquet, O., Simon-Gabriel, C.-J., and Schölkopf, B. (2017). AdaGAN: Boosting Generative Models. *arXiv preprint arXiv:1701.02386*.

[Tong et al., 2017] Tong, T., Ledig, C., Guerrero, R., Schuh, A., Koikkalainen, J., Tolonen, A., Rhodius, H., Barkhof, F., Tijms, B., Lemstra, A. W., Soininen, H., Remes, A. M., Waldemar, G., Hasselbalch, S., Mecocci, P., Baroni, M., Lötjönen, J., van der Flier, W., and Rueckert, D. (2017). Five-class differential diagnostics of neurodegenerative diseases using random undersampling boosting. *NeuroImage: Clinical*, 15:613–624.

[Tsunoda et al., 2014] Tsunoda, Y., Moribe, M., Orii, H., Kawano, H., and Maeda, H. (2014). Pseudo-normal image synthesis from chest radiograph database for lung nodule detection. In *Advances in Intelligent Systems and Computing*, volume 268, pages 147–155. Springer.

[Tustison et al., 2010] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320.

[Udrea and Mitra, 2017] Udrea, A. and Mitra, G. D. (2017). Generative Adversarial Neural Networks for Pigmented and Non-Pigmented Skin Lesions Detection in Clinical Images. In

Proceedings - 2017 21st International Conference on Control Systems and Computer, CSCS 2017, pages 364–368. IEEE.

[Valdes Hernandez et al., 2015] Valdes Hernandez, M. D. C., Armitage, P. A., Thrippleton, M. J., Chappell, F., Sandeman, E., Munoz Maniega, S., Shuler, K., and Wardlaw, J. M. (2015). 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):e00415.

[Valdés Hernández et al., 2013] Valdés Hernández, M. d. C., Morris, Z., Dickie, D. A., Royle, N. A., Muñoz Maniega, S., Aribisala, B. S., Bastin, M. E., Deary, I. J., and Wardlaw, J. M. (2013). Close Correlation between Quantitative and Qualitative Assessments of White Matter Lesions. *Neuroepidemiology*, 40(1):13–22.

[van Dijk et al., 2008] van Dijk, E. J., Prins, N. D., Vrooman, H. a., Hofman, A., Koudstaal, P. J., and Breteler, M. M. B. (2008). Progression of cerebral small vessel disease in relation to risk factors and cognitive consequences: Rotterdam Scan study. *Stroke; a journal of cerebral circulation*, 39(10):2712–9.

[Van Nguyen et al., 2015] Van Nguyen, H., Zhou, K., and Vemulapalli, R. (2015). Cross-domain synthesis of medical images using efficient location-sensitive deep network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 677–684. Springer.

[Vemulapalli et al., 2015] Vemulapalli, R., Nguyen, H. V., and Zhou, S. K. (2015). Unsupervised Cross-modal Synthesis of Subject-specific Scans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 630–638.

[Vemuri et al., 2011] Vemuri, P., Simon, G., Kantarci, K., Whitwell, J. L., Senjem, M. L., Przybelski, S. A., Gunter, J. L., Josephs, K. A., Knopman, D. S., Boeve, B. F., Ferman, T. J., Dickson, D. W., Parisi, J. E., Petersen, R. C., and Jack, C. R. (2011). Antemortem differential diagnosis of dementia pathology using structural MRI: Differential-STAND. *NeuroImage*, 55(2):522–531.

[Vincent et al., 2010] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408.

[Wang et al., 2017] Wang, D., Gu, C., Wu, K., and Guan, X. (2017). Adversarial neural networks for basal membrane segmentation of microinvasive cervix carcinoma in histopathology images. In *Proceedings of 2017 International Conference on Machine Learning and Cybernetics, ICMLC 2017*, volume 2, pages 385–389. IEEE.

[Wang et al., 2018] Wang, Y., Yu, B., Wang, L., Zu, C., Lalush, D. S., Lin, W., Wu, X., Zhou, J., Shen, D., and Zhou, L. (2018). 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. *NeuroImage*, 174:550–562.

[Wang and Bovik, 2002] Wang, Z. and Bovik, A. C. (2002). A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84.

[Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

[Wardlaw et al., 2013] Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R. I., O'Brien, J. T., Barkhof, F., Benavente, O. R., Black, S. E., Brayne, C., Breteler, M., Chabriat, H., DeCarli, C., de Leeuw, F. E., Doubal, F., Duering, M., Fox, N. C., Greenberg, S., Hachinski, V., Kilimann, I., Mok, V., van Oostenbrugge, R., Pantoni, L., Speck, O., Stephan, B. C., Teipel, S., Viswanathan, A., Werring, D., Chen, C., Smith, C., van Buchem, M., Norrving, B., Gorelick, P. B., and Dichgans, M. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12(8):822–838.

[Warren et al., 2013] Warren, J. D., Rohrer, J. D., and Rossor, M. N. (2013). Clinical review. Frontotemporal dementia. *BMJ (Clinical research ed.)*, 347(aug12 3):f4827–f4827.

[Weiner, 2014] Weiner, M. W. (2014). Alzheimer's Disease Neuroimaging Initiative 2 (ADNI2) Protocol (ADC-039). `http://adni.loni.usc.edu/wp-content/themes/freshnews-dev-v2/documents/clinical/ADNI-2_Protocol.pdf`.

[Weiner et al., 2013] Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., et al. (2013). The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111–e194.

[Weiner et al., 2017] Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Morris, J. C., Petersen, R. C., Salazar, J., Saykin, A. J., Shaw, L. M., Toga, A. W., and Trojanowski, J. Q. (2017). The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement. *Alzheimer's and Dementia*, 13(5):561–571.

[White, 2016] White, T. (2016). Sampling generative networks. *arXiv preprint arXiv:1609.04468*.

[Wold et al., 1987] Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

[Wolterink et al., 2017] Wolterink, J., Leiner, T., Viergever, M., and Isgum, I. (2017). Generative adversarial networks for noise reduction in low-dose CT. *IEEE Transactions on Medical Imaging*, 36(12):2536–2545.

[Wu et al., 2016] Wu, J., Zhang, C., Xue, T., Freeman, B., and Tenenbaum, J. (2016). Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, pages 82–90.

[Wu et al., 2017] Wu, X., Xu, K., and Hall, P. (2017). A Survey of Image Synthesis and Editing with Generative Adversarial Networks. *Tsinghua Science & Technology,*, 22(6):660–674.

[Xue et al., 2017] Xue, Y., Xu, T., Zhang, H., Long, R., and Huang, X. (2017). Segan: Adversarial network with multi-scale l_1 loss for medical image segmentation. *Neuroinformatics*, pages 1–10.

[Yan et al., 2018] Yan, P., Xu, S., Rastinehad, A. R., and Wood, B. J. (2018). Adversarial Image Registration with Application for MR and TRUS Image Fusion. *arXiv preprint arXiv:1804.11024*.

[Yang et al., 2017] Yang, D., Xu, D., Zhou, S. K., Georgescu, B., Chen, M., Grbic, S., Metaxas, D., and Comaniciu, D. (2017). Automatic Liver Segmentation Using an Adversarial Image-to-Image Network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 507–515. Springer.

[Ye et al., 2013] Ye, D. H., Zikic, D., Glocker, B., Criminisi, A., and Konukoglu, E. (2013). Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 606–613. Springer.

[Yeh et al., 2016] Yeh, R. A., Chen, C., Lim, T. Y., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. (2016). Semantic Image Inpainting with Deep Generative Models. *arXiv preprint arXiv:1607.07539*.

[Yoo et al., 2016] Yoo, D., Kim, N., Park, S., Paek, A. S., and Kweon, I. S. (2016). Pixel-level domain transfer. In *European Conference on Computer Vision*, pages 517–532. Springer, Cham.

[Zhang et al., 2017] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. (2017). StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:5908–5916.

[Zhang et al., 2018] Zhang, J., Jia, K., Jia, J., and Qian, Y. (2018). An improved approach to infer protein-protein interaction based on a hierarchical vector space model. *BMC Bioinformatics*, 19(1).

[Zhang et al., 2001] Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57.

[Zhang et al., 2012] Zhang, Y., Wu, G., Yap, P. T., Feng, Q., Lian, J., Chen, W., and Shen, D. (2012). Hierarchical patch-based sparse representation-a new approach for resolution enhancement of 4D-CT lung data. *IEEE Transactions on Medical Imaging*, 31(11):1993–2005.

[Zhao et al., 2016] Zhao, J., Mathieu, M., and LeCun, Y. (2016). Energy-based Generative Adversarial Network. *arXiv preprint arXiv:1609.03126*.

[Zhu et al., 2017a] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017a). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

[Zhu et al., 2018] Zhu, X., Liu, Y., Li, J., Wan, T., and Qin, Z. (2018). Emotion classification with data augmentation using generative adversarial networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 349–360. Springer, Cham.

[Zhu et al., 2017b] Zhu, X., Liu, Y., Qin, Z., and Li, J. (2017b). Data Augmentation in Emotion Classification Using Generative Adversarial Networks. *arXiv preprint arXiv:1711.00648*.

Appendix A

Dementia Diagnosis

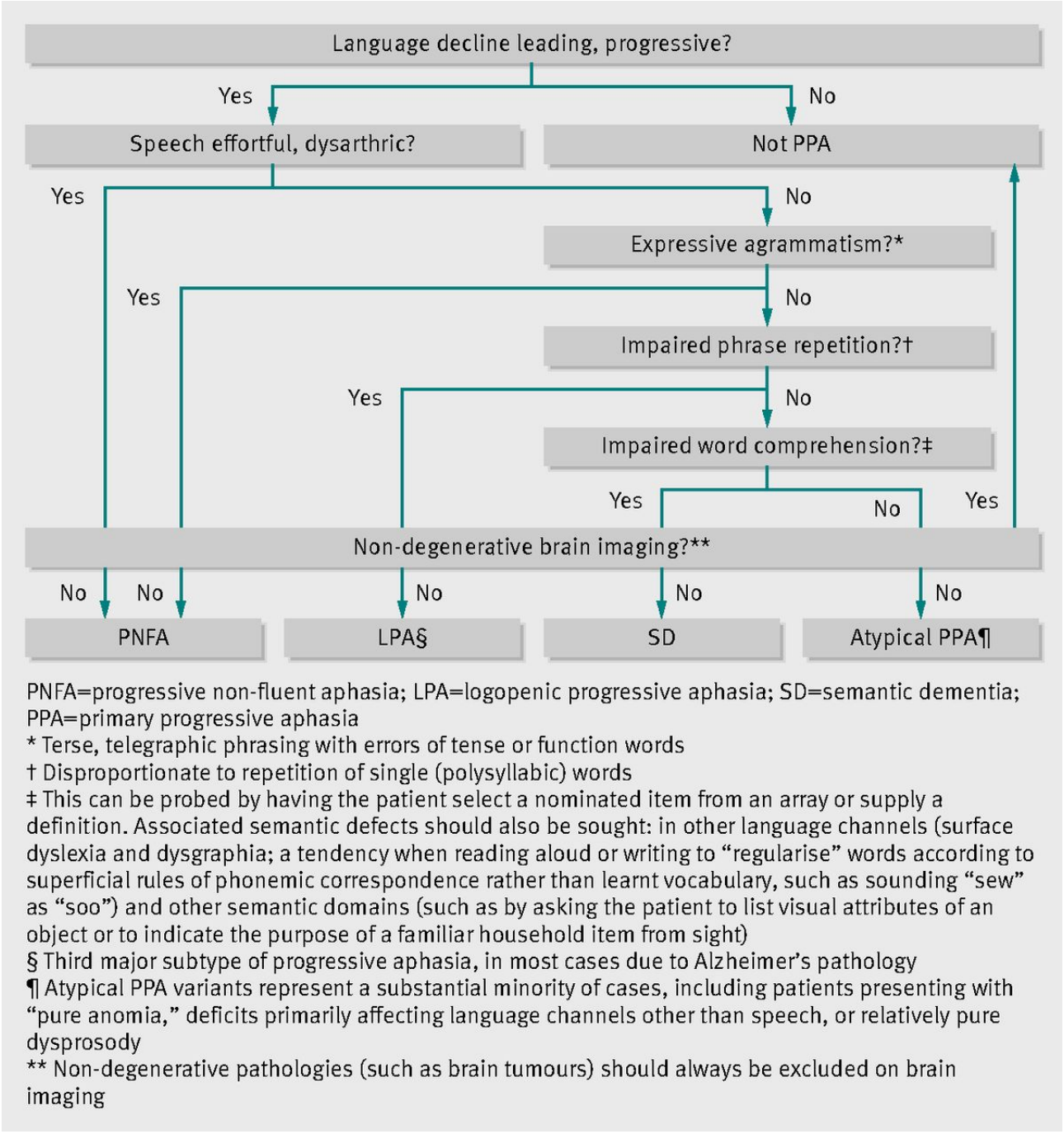


Figure A.1: Bedside clinical assessment of the progressive aphasias: a simple algorithm (informed by current consensus criteria for progressive aphasia6) for syndromic diagnosis of patients presenting with progressive language decline. The clinical syndromic diagnosis should be supplemented by neuropsychological assessment, brain magnetic resonance imaging, and ancillary investigations including cerebrospinal fluid examination. *Warren, J. D., Rohrer, J. D., & Rossor, M. N. (2013). Frontotemporal dementia. BMJ, 347(aug12 3), f4827f4827. http://doi.org/10.1136/bmj.f4827*

Main behavioural features and subtypes*	Examples
Disinhibition:	
Socially inappropriate behaviour	Inappropriately approaching, touching, or kissing strangers, verbal or physical aggression, fatuity, staring
Loss of manners or decorum	Inappropriate laughter, jokes, or opinions that may be offensive to others, faux pas, lack of etiquette, altered dress sense
Impulsive, rash, or careless actions	Reckless driving, new onset gambling, buying or selling objects without regard for consequences
Apathy and inertia:	
Apathy	Lacking initiative, ceasing to engage in former activities or hobbies, poor personal hygiene
Inertia	Needs prompting to initiate or continue routine activities, less likely to initiate or sustain a conversation
Reduced autonomy	Environmental dependency, utilisation behaviours (such as handling or using items or reading signs aloud when not required or appropriate to social context)
Loss of sympathy and empathy:	

Diminished response to other peoples needs and feelings	Making hurtful comments or disregarding other peoples pain or distress, less warmth or interest toward others (such as grandchildren, pets), hypoemotionality, failure to appreciate ambiguous social signals (such as sarcasm)
Diminished social interest, interrelatedness, or personal warmth	Decrease in social engagement, emotional detachment, distant from friends and relatives, reduced libido, altered sense of humour
Perseverative, stereotyped, and compulsive or ritualistic behaviour:	
Simple repetitive movements	Tapping, clapping, rubbing, scratching, picking at self, humming, rocking
Complex, compulsive, or ritualistic behaviours	Counting and cleaning rituals, collecting or hoarding, checking, ordering objects, walking fixed routes, clock watching, new obsessional interests or preoccupations (such as religiosity, musicophilia)
Stereotypy of speech	Habitual repetition of words, phrases, or themes
Hyperorality and dietary changes:	
Altered food preferences	Carbohydrate cravings (particularly sweets), food fads
Binge eating, increased consumption of alcohol or cigarettes	Consuming excessive amounts of food, gluttony, rapid, messy eating, overfilling mouth, compulsive use of alcohol or smoking

Oral exploration or consumption of inedible objects	Pica, features of Kluver-Bucy syndrome
Loss of insight	Unaware of or unconcerned by difficulties
Others:	
Psychotic features	Hallucinations (especially somatic or visual), delusions (especially somatic or paranoid)
Altered sensitivity to pain	Hypochondriasis, heightened distress with innocuous stimuli, lack of distress in response to painful stimuli
Altered temperature sensibility	Dressing inappropriate to climate

Table A.1: *Within the broad phenotype of behavioural variant frontotemporal dementia; clinical features in individual patients are highly variable. Early features are often loss of warmth and empathy, social faux pas, and altered eating behaviour or food preferences. Especially in association with expansions in the C9ORF72 gene. *Warren, J. D., Rohrer, J. D., & Rossor, M. N. (2013). Frontotemporal dementia. BMJ, 347(aug12 3), f4827f4827. http://doi.org/10.1136/bmj.f4827*

Appendix B

Summary table of patch based synthesis methods

Method	Library $\{A, A'\}$	Search solution	Speed solution	Coherence solution	Normalisation solution	Application
Image Analogies [Hertzmann et al., 2001]	{Single source image, Single target image}	l_2-norm + coherence metric	Small library, approximate nearest neighbour	Augmented similarity metric	Match mean and standard deviation	Image filters, texture synthesis, super resolution, texture transfer, artistic filters, texture by numbers
[Roy et al., 2014b, Roy et al., 2010]	{Single T_1 and T_2 image pair, Single FLAIR image}	l_2-norm + coherence metric	Small library	Augmented similarity metric, non-local means patch combination	Assumed	FLAIR synthesis, super-resolution
MIMECS [Roy et al., 2011, Roy et al., 2013]	Varies, {Single 2D image, single MR image}	Compressed sensing	Restricted search space from tissue segmentations	NA	Set peak of WM histogram to 1	Longitudinal data normalisation, atlas construction, contrast normalisation, distortion correction in diffusion images, super resolution and FLAIR synthesis.

[Iglesias et al., 2013]	{Multiple PD images, Multiple T_1 images}	l_2-norm	Restricted search space to local area, multi-resolution	None	FreeSurfer tool, no details	Registration, segmentation
[Zhang et al., 2012]	{Low res 4D-CT images, High res 4D-CT images}	l_2-norm + coherence measure, sparsity constraint	Restricted search to patches taken from expected anatomical location at different time point, cropped images	Augmented similarity metric	None mentioned, CT quantitative	CT super resolution
[Rueda et al., 2013]	{Low res MR images, High res MR images}	l_2-norm, sparsity constraint	Restricted search space from tissue segmentations	Post-processing regularisation through back projection	None mentioned	Super resolution

Modality Propagation [Ye et al., 2013]	Varies, {Multiple MR images, Multiple 2D images}	l_2-norm (1^{st} iteration), l_2-norm + coherence metric (subsequent iterations)	Restricted search space to local area and locally similar images	Iterative refinement	Histogram matching	DTI-FA + T_2 synthesis from T_1, Pseudo-healthy T_2 synthesis
[Tsunoda et al., 2014]	{Multiple non-pathological chest radiographs}	Normalised correlation coefficient	Restricted search space to local area, down sampling images	None	NA (use of similarity metric)	Pseudo-healthy radiograph synthesis
[Cao et al., 2013]	Varies, {Multiple microscopy images, Multiple microscopy images}	EM, sparse coding	Small library	None	None mentioned	Multi-modal registration, texture synthesis
[Roy et al., 2014a]	Varies, {Single MR image, Single CT image}	EM	Restricted search by finding a set of similar patches	None	None mentioned	MR to CT registration

[Dawant et al., 2012, Cao et al., 2014]	Varies, {Single microscopy image, Single microscopy image}	Sparse coding, l_2-norm	Small library	None	Scaling to $[0, 1]$	Multi-modal registration, texture synthesis

Table B.1: Comparison of image synthesis methods based loosely on the Image Analogies [Hertzmann et al., 2001] framework. This comparison includes: A) The method name (if provided). B) What image modalities were used as the source and target modalities. C) The solution to finding the nearest patch. D) The solution to the problem of search speed. E) The solution to the issue of producing a visually coherent image from distinct patches. F) The solution to the issue of intensity normalisation. G) The application the method was used for.

Appendix C

StitchGAN: Generating high resolution 2/3D images and surface data using generative adversarial networks

C.1 Introduction

The level of detail required in modern day medical diagnostics continues to push images to higher and higher resolutions. To handle this high resolution, computer vision techniques used in the analysis of such images are often patch based, multi-resolution, or rely on dimensionality reducing pre-processing steps. This avoids the need to process the entire image at once, reducing the computational complexity of these methods to manageable levels. Part of the enormous success of Convolutional Neural Networks (CNNs) in this field can be attributed to their convolutional nature making many architectures almost invariant to image size. Those architectures which are often used to process a whole image, such as UNet [Ronneberger et al., 2015] and auto-encoders [Vincent et al., 2010], can often benefit from pretraining [Sevetlidis et al., 2016] or scale well with image size.

However, Generative Adversarial Networks (GANs) are notoriously unstable and have long

training times even at modest resolutions. Because of this, GANs do not extend well to higher dimensions and larger image sizes. GANs have been shown to have many exciting potential applications in medical imaging, however this major hurdle must first be overcome. In this section we propose an alternative to ever increasing memory and GPU requirements by generating large images through stitching the output of multiple small GANs together in a locally and globally consistent way.

C.1.1 Related work

In [Yeh et al., 2016], the authors use GANs to perform image inpainting on a number of datasets. In their experiments, the authors remove up to 80% of pixels from an image and imputes their values by using a GAN to generate an image which is consistent with the remaining pixels. This is done by performing a gradient descent over the generator input $\mathbf{z}$, finding the $\hat{\mathbf{z}}$ which minimises a cost function which penalises a weighted L_1-norm between the generated image and the remaining pixels, and a high discriminator loss.

C.1.2 Contribution

We propose StitchGAN, a method of generating high resolution 2-dimensional (2D) and 3-dimensional (3D) images without increasing the memory or computational requirements, allowing for arbitrarily large images be generated, beyond that which might usually fit in GPU memory. We demonstrate that the approach also has the advantage of not being constrained to simple 2/3D images, and can also be used to generate surface data.

C.2 Method

In StitchGAN, we use multiple independently trained GANs to generate images from different locations within a larger image. The ultimate goal is to generate high resolution 2D and 3D images by joining together these lower resolution sub-images. Naively joining the output from

multiple GANs would be insufficient, as, even if the sub-images are completely realistic, the resulting image will be disjointed and unrealistic. This presents two problems: how to ensure that each sub-image corresponds to the same image (global coherency), and how to make sure the joins between the sub-images are not visible (local coherency). We first describe the method in terms of a 2D image, before later showing how it can be adapted to generate 3D images and surface data.

C.2.1 Global coherency

We adapt the approach of [Yeh et al., 2016] by using a two-times under-sampled base image as a foundation for the full resolution image. The pixels from this low resolution image are then re-distributed throughout the higher resolution image space, providing every second pixel of a high resolution image. The problem can then be described as one of missing information, and the inpainting approach of [Yeh et al., 2016] employed. Since the image is now four-times as large, four GANs are trained, one to perform inpainting in each quadrant of the high resolution image.

Five GANs are therefore trained. The first is trained on images formed by taking every second pixel along each dimension of the training images, and is then used to generate the synthetic undersampled base image. The remaining four are trained on sub-images covering each quadrant of the training images.

Sub-images which are consistent with the base image are generated by finding the optimal $\hat{\mathbf{z_k}}$ for each quadrant k using:

$$\hat{\mathbf{z}}_\mathbf{k} = \arg\min_{\mathbf{z}} \{\zeta \left\| L(G_k(\mathbf{z})) - x_k \right\|^1 + \lambda D_k(G_k(\mathbf{z}))\} \tag{C.1}$$

where x_k represents the region in the low resolution image which corresponds to generator G_k and associated discriminator D_k, L is a downsampling operator which samples every second pixel and λ and ζ are weighting factors which control the influence of the two terms. The first

term ensures that every second pixel in the generated sub-image matches the corresponding pixel in the base image. The second term ensures that the generated image is still considered real by the discriminator, and therefore remains realistic.

C.2.2 Local coherency

While Equation C.1 ensures that each sub-image is consistent with a low resolution image, there will be new high frequency features generated within each sub-image. If these features extend to the edge of a sub-image there will be discontinuities in the joint image if they are not continued by their neighbours. To account for this we use overlapping sub-images and add a third term with associated weight γ to Equation C.1:

$$\begin{aligned}\hat{\mathbf{z}_\mathbf{k}} = \arg\min_{\mathbf{z}} \{\zeta \left\| L(G_k(\mathbf{z})) - x_k \right\|_M^1 + \lambda D_k(G_k(\mathbf{z})) + \\ \gamma \left\| (G_k(\mathbf{z}) - y_k) \right\|_M^1 \}\end{aligned} \tag{C.2}$$

where $\mathbf{y}$ is an image containing any overlap from the sub-image's neighbours which have already been generated, M is a binary mask indicating these regions, and $\|\|_M$ signifies the norm evaluated only over the region indicated by M.

To generate a full high resolution image, a low resolution base image is first generated. Next, a high resolution seed sub-image is generated at one location consistent with the base image using Equation C.1. Starting from the seed sub-image's neighbours, the remaining sub-images are then generated using Equation C.2 so as to be consistent with both the base image and any high resolution neighbours which have already been generated. Finally, each sub-image is blended with its neighbours across a linear gradient. The training and inference procedures and equations are shown graphically in Figures C.1 and C.2.

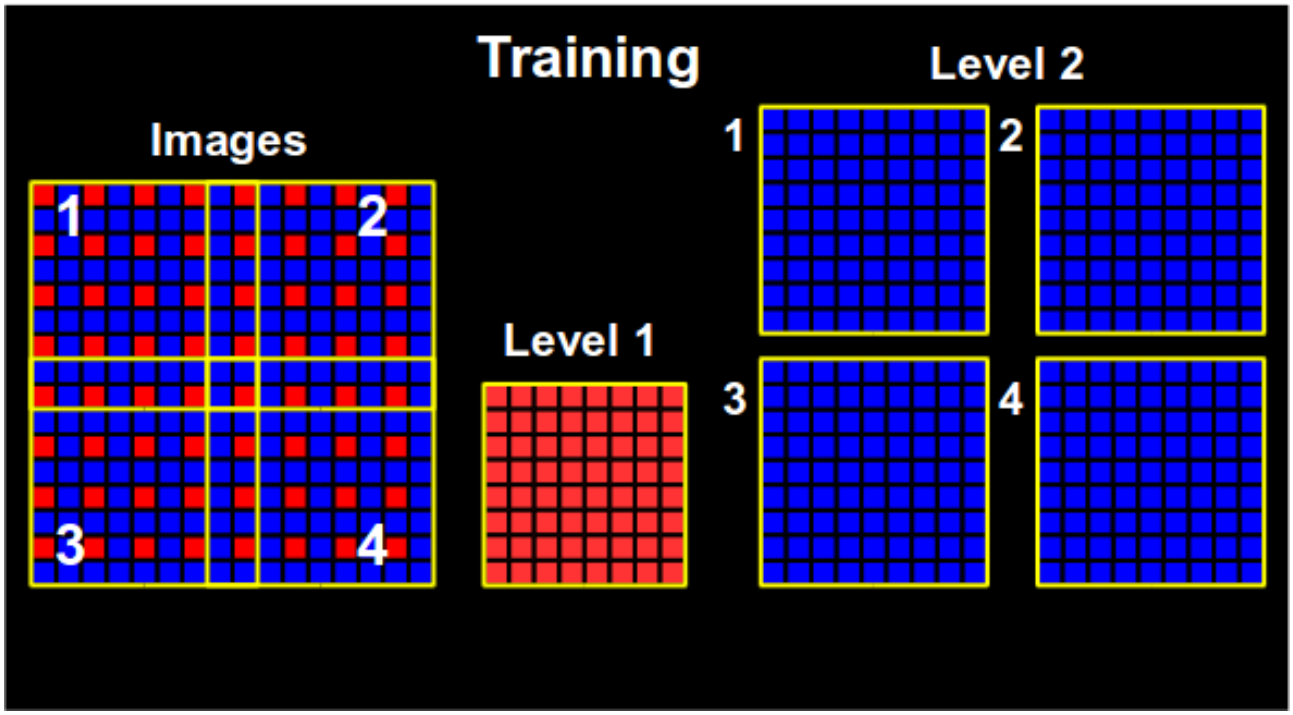


Figure C.1: Division of a full resolution 2D image into a low resolution base image (Level 1) and 4 overlapping sub images (Level 2). A single GAN is trained on each set of images.

C.2.3 Recursive generation

The proposed method is recursive, in that once a higher resolution image has been generated, it can form the base image for an even higher resolution. In this way, arbitrarily high resolution images can be generated. However, the number of networks required increases as the square (2D) or cube (3D) of the number of resolution levels, so generating images more than 4-8 times the size of that produced by an individual network becomes computationally challenging. In each of the experiments shown here we use three resolution levels.

C.2.4 Similarities to dictionary learning

The proposed method can also be thought of an extension to patch based dictionary learning methods such as those discussed in Section 3.1.3, whereby images are constructed by selecting patches from a stored dictionary which best fit a set of criteria, often local and global coherency. Within this paradigm, the dictionary is replaced with a GAN, allowing us to sample from the entire learned manifold of potential patches, as opposed to the discrete points stored in a

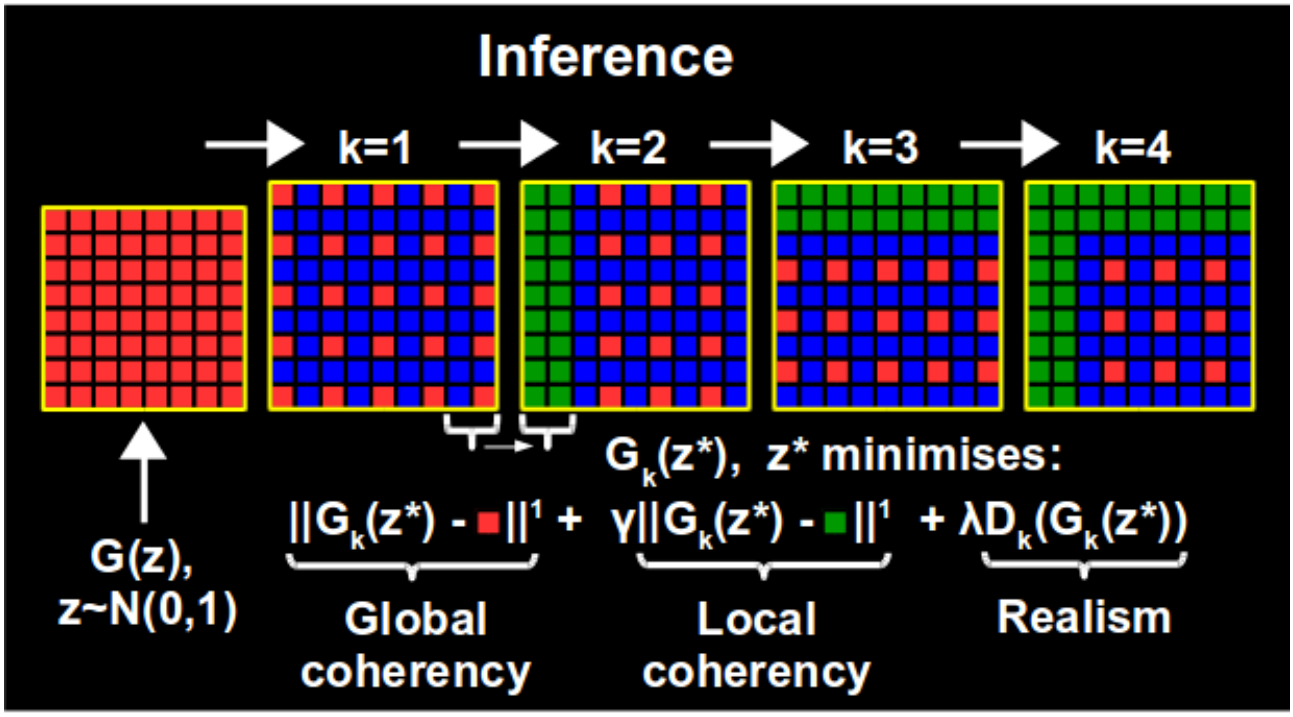


Figure C.2: Proposed inference procedure for a 2D image. A low resolution base image is first generated (red). The generated pixels are then distributed throughout the 4 sub-images ($k = 1, 2, 3, 4$). Each sub image is generated in turn following the given equation to ensure consistency with the distributed base image pixels (red) and overlapping regions from previously generated sub-images (green).

dictionary. This solves the search speed problem discussed i Section 3.1.3 by providing an efficient GPU based gradient descent approach to patch matching, and addresses the problem of enforcing local coherency by including it directly in the loss function.

C.2.5 Dynamic balancing of the generator and discriminator

One important, but rarely discussed, hyperparameter for GAN training is the ratio between the discriminator (D) and generator (G) update cycles (referred to as k in [Goodfellow et al., 2014]). Setting $k = 1$ is a common choice, however some propose more complex strategies, often with little justification. For example, the authors of [Arjovsky et al., 2017] opt for a scheme with $k = 5$ during normal training an $k = 100$ at the start of training and briefly for pre-defined intervals thereafter. [Goodfellow et al., 2014] present the choice of k as a balance between ensuring the optimality of D (perfect real / synthetic discrimination, a desired property) and avoiding unnecessary and time consuming update cycles, and overfitting, when D is already at

or near optimality. Intuitively, an optimal D allows for the most useful information to be fed back to G, while a completely deficient D will provide nothing useful for G. We therefore have two goals to maximise training efficiency: First, to only update G when D is near optimal, and second, to only update D when D is far from optimal. With this in mind we propose the following simple heuristic to dynamically adjust the ratio between D and G update cycles to maximise the useful information passed to G, while minimising wasted training cycles.

while *Stopping criteria not reached* **do**
  repeat
    Sample batch m of size n from noise distribution
    Compute $E^G = \mathrm{D}(\mathrm{G}(m))$
    Compute sensitivity as $\frac{1}{n}\sum E^G > 0$
    Sample batch r of size n from real images
    Compute $E^R = \mathrm{D}(r)$
    Compute specificity as $\frac{1}{n}\sum E^R < 0$
    Update D
  until *specificity* > 0.75 *and sensitivity* > 0.75;
  repeat
    Sample batch m of size n from noise distribution
    Compute $E^G = \mathrm{D}(\mathrm{G}(m))$
    Compute sensitivity as $\frac{1}{n}\sum E^G > 0$
    Update G
  until *sensitivity* < 0.25;
end

Algorithm 3: Proposed dynamic training schema.

As well as speeding up and stabilising training, the proposed schema provides some useful insight into the health of the system during training. Long periods of generator training (within the second inner loop of 3) proved to be indicative of impending mode collapse. The most stable training was associated with frequent swaps (no more than 50 iterations in one loop) between the two inner loops, smooth changes in sensitivity and specificity, and an average of 5-10 times more G updates than D updates.

C.3 Experiments

To demonstrate some potential applications of StitchGAN, we apply the technique to three typical types of neurological imaging data. First, we demonstrate the method using simple

2D slices from a T_1-weighted 2D image. Despite these being within the abilities of of some of the larger GAN formulations, we include these experiments for demonstration purposes. Next, we show how the method can be used to generate cortical surfaces, before finally applying the technique in 3D to generate 2-mm isotropic resolution 3D volumes.

C.3.1 2D Slice

In this experiment we generate 1mm isotropic 2D slices using data from Alzheimers Disease Neuroimaging Initiative (ADNI) [Mueller et al., 2005]. All images collected as part of the Alzheimers Disease Neuroimaging Initiative - 2 (ADNI-2) study were used, providing 6200 training images. Data was bias corrected (N4 [Tustison et al., 2010]), skull stripped (PINCRAM [Heckemann et al., 2015]) and rigidly co-registered to a 1mm isotropic space. A single radial slice was selected and used for training. To create the initial base image training set, each image was down-sampled to a 4mm resolution giving an image size of 48-by-48px. For the sub-image training set, 4 sets of overlapping 48-by-48px sub-images were extracted from 2mm resolution down-sampled images, and 16 sets extracted from the original images (Figure C.3). Due to the symmetry of the brain, the total number of sub-image sets can be halved by grouping corresponding sub-images from opposite hemispheres together, leaving 11 GANs to be trained.

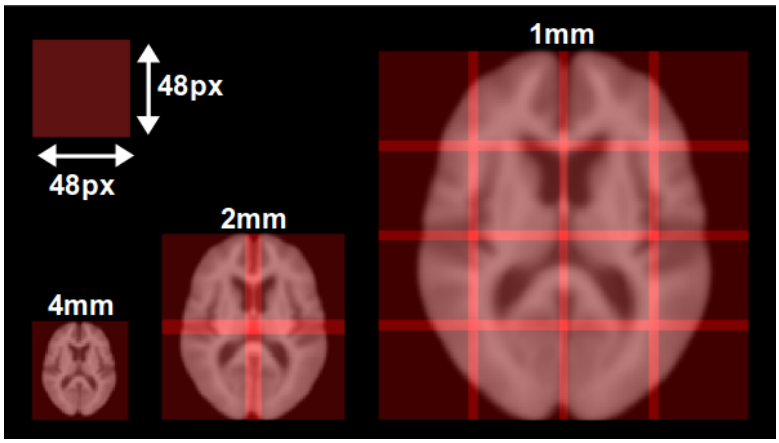


Figure C.3: Parcellation of 1mm isotropic 2D slice into a base image and overlapping sub-images at different resolutions. Red squares indicate locations of each image set overlayed on the average training image.

C.3.2 Spherical surface projection of sulcal depth

An advantage of the proposed method is that generated images need not be simple planes. Here we demonstrate how a surface describing the shape of an individual's brain can be generated. Data for this experiment comes from 618 subjects from the Human Connectome Project [1], which provides pre-computed sulcal depth maps and surface templates. Figure C.4 shows an example of a typical sulcal depth map used to a displace an average "very-inflated" cortical surface template, projected onto a sphere, and projected onto a 2D plane using Mollweide projection (as in [Kang et al., 2012].

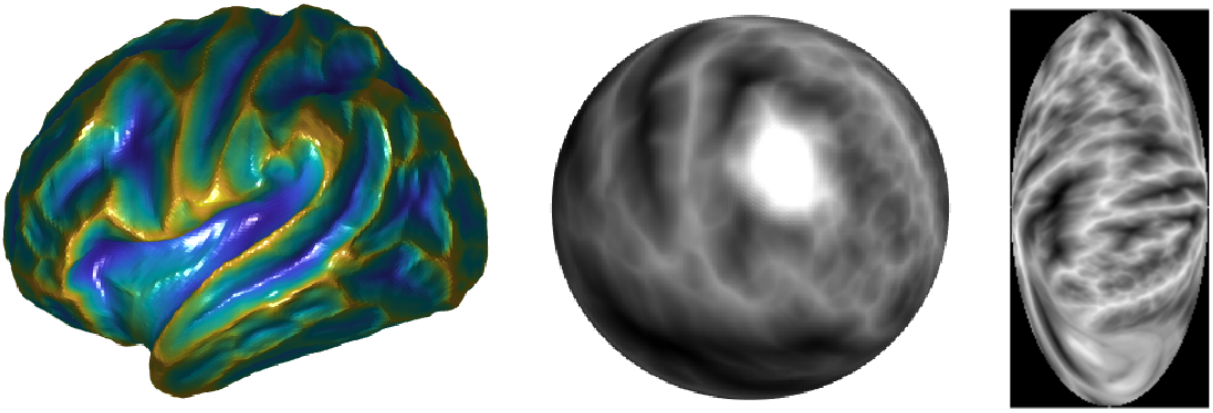


Figure C.4: Example cortical surface map in three spaces. Left: Values used to displace corresponding vertices on an average "very-inflated" cortical surface template. Middle: Projected onto a sphere. Right: Projected onto a 2D plane using Mollweide projection.

A sulcal depth map describes the displacement between an "inflated" and "very-inflated" White Matter (WM) surface. By generating plausible synthetic sulcal depth maps and using these to displace an average "very-inflated" surface, a set of synthetic cortical surfaces can be generated.

The synthetic depth maps are generated piece-wise on the spherical projection using SitchGAN. A 128-by-64px Mollweide projection forms the $\frac{1}{4}$ resolution base image, with 6 overlapping 96-by-96px sub images covering the $\frac{1}{2}$ resolution image, and another 24 covering the full resolution image (Figure C.5).

As well as being able to generate high resolution depth maps, the proposed method also has

[1] http://www.humanconnectomeproject.org/

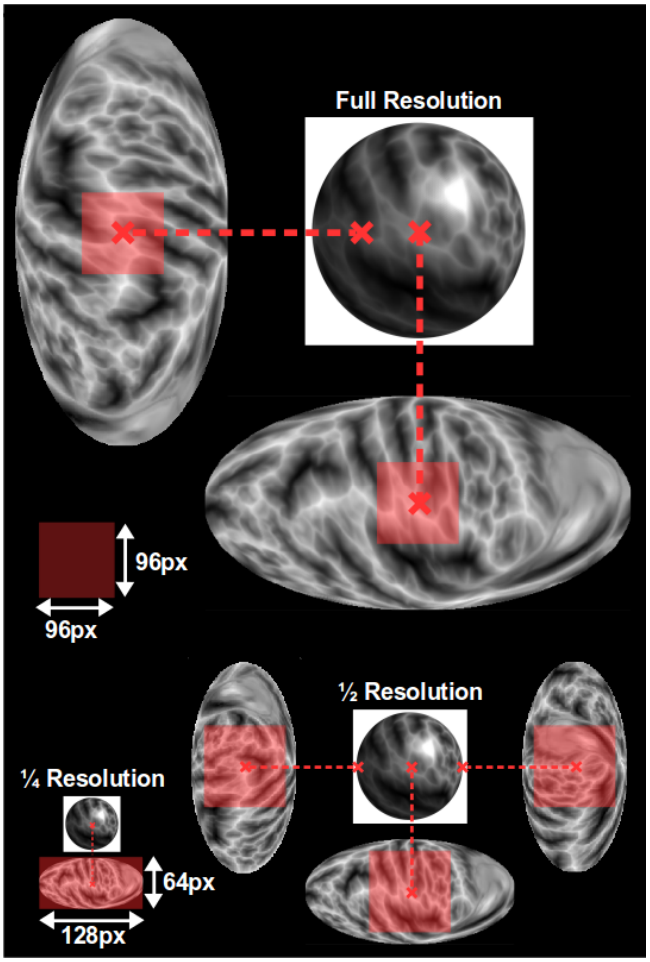


Figure C.5: Parcellation of spherically projected sulcal depth map into a base image and overlapping sub-images at different resolutions. Red squares indicate the location of the sub-images on the 2D Mollweide projection. Only 2/24 locations shown at full resolution, and 3/6 at $\frac{1}{2}$ resolution.

the advantage of only generating small regions from the surface at a time, and is therefore significantly less effected by projection artefacts than if the whole surface was generated at once.

C.3.3 3D volume

Using the same sample of ADNI data as used previously, we now generate full 3D volumes. An optimal parcellation for highest resolution sub-images was formed so as to minimise the number of individual networks required whilst maintaining a minimum overlap of 4mm (Figure C.6).

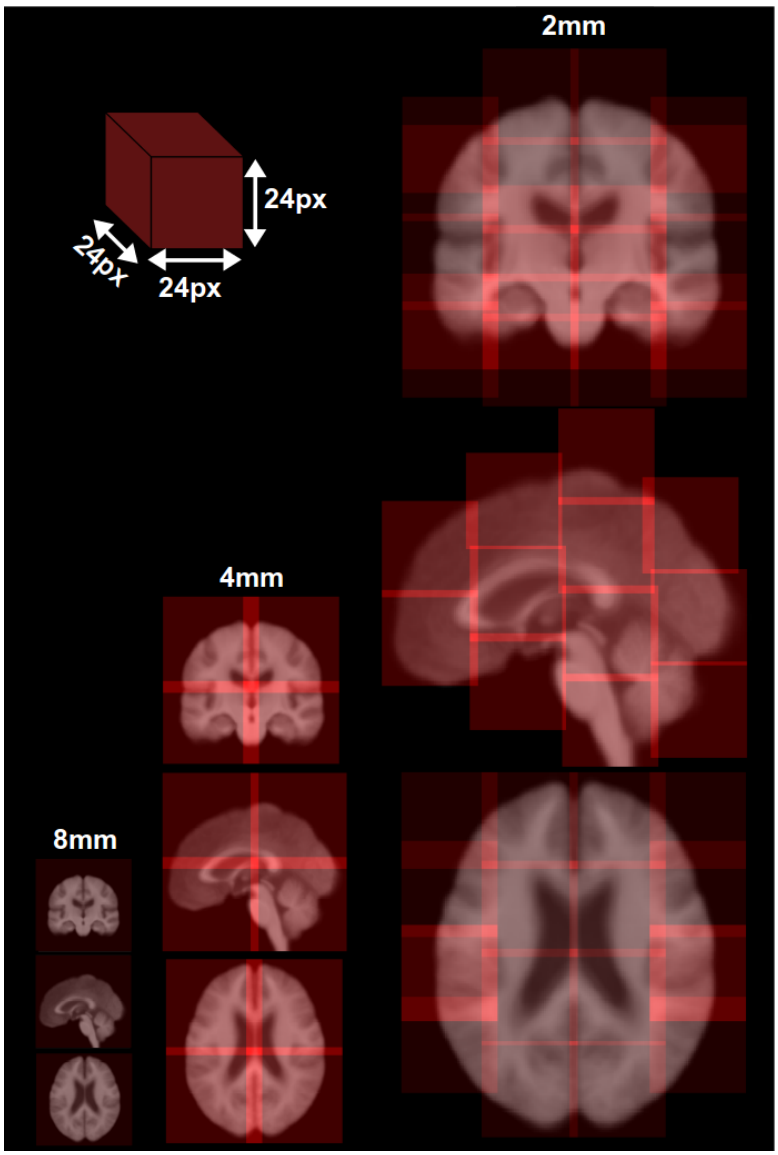


Figure C.6: Parcellation of 2mm isotropic 2D volume into a base image and overlapping sub-images at different resolutions. Red squares indicate locations of each image set overlayed on the average training image.

C.3.4 Network architecture and parameters

The Wasserstein Generative Adversarial Network (WGAN) [Arjovsky et al., 2017] architecture was used for all experiments. The architecture of each network was adjusted to fit the required resolution of the experiment. For the 3D networks, the 2D convolutions were changed to 3D. All experiments used a batch size of 64, generator input size of 100, and generator and discriminator learning rates of 5×10^{-5}. Batch normalisation was used in the discriminator only and optimisation was performed using stochastic gradient descent.

When generating sub-images, the first is generated using Equation C.1, with subsequent images generated using Equation C.2. Parameters ζ, λ, γ and learning rate α were chosen empirically for each dataset and resolution level by examining the generated images. ζ controls how strongly consistency with the previous resolution level was enforced. It was found to be important for the 2D applications, but less so in 3D. This can be attributed to the 3D patches covering a smaller anatomical region, with less potential for anatomical variation, in addition to the more complete contextual information provided by potential overlaps in the additional dimension. Consistency with surrounding patches was therefore seen to be sufficient produce realistic results, with a small ζ value only required to guide the process in cases with little or no overlap with existing patches. λ controls the influence of the discriminator loss on the generated patches. In [Yeh et al., 2016], this was necessary to ensure only realistic faces were proposed. In our application, we found that there was little observable difference between the realism of patches scored highly/lowly by discriminator at the end of training, and therefore a high λ value only lead to less consistent images. This could be for a few reasons. Firstly the manifolds of image patches used in these experiments are likely simpler than the manifold of faces used by [Yeh et al., 2016]. The more complex manifolds are more difficult to learn and could therefore lead to regions which do not map to realistic faces, requiring the λ term to encourage solutions away from these regions. The simpler manifolds are easier to learn without such regions, and therefore do no not require the λ term. The other possibility could be the use of WGAN in place of the Deep Convolutional Generative Adversarial Network (DCGAN) used in [Yeh et al., 2016], potentially leading to higher quality manifolds being learned. λ was therefore set to 0 in all 2D experiments, and to 0.025 in the 3D experiments. This small value was found to be necessary as the learned manifolds for the more complex 3D patches could occasionally produce unrealistic images. λ, controlling local coherency, was kept at 1 in all experiments except when generating the highest resolution 2D slice, where it was reduced to 0.2. In this case, a high value was not necessary to enforce local coherency, however reducing it would lead to greater global coherency, particularly in the form of symmetry. Finally the value of α showed little influence on the quality of the final images, but was tuned for each set of experiments so as to ensure convergence in a short a time as possible.

C.4 Results

Sample generated images showing 2D, surface and 3D images, along with reference real images, can be seen in Figures C.8, C.9 and C.7. Images at each resolution level are shown to demonstrate the progression from under sampled base image to final full resolution image, as well as exemplar real images.

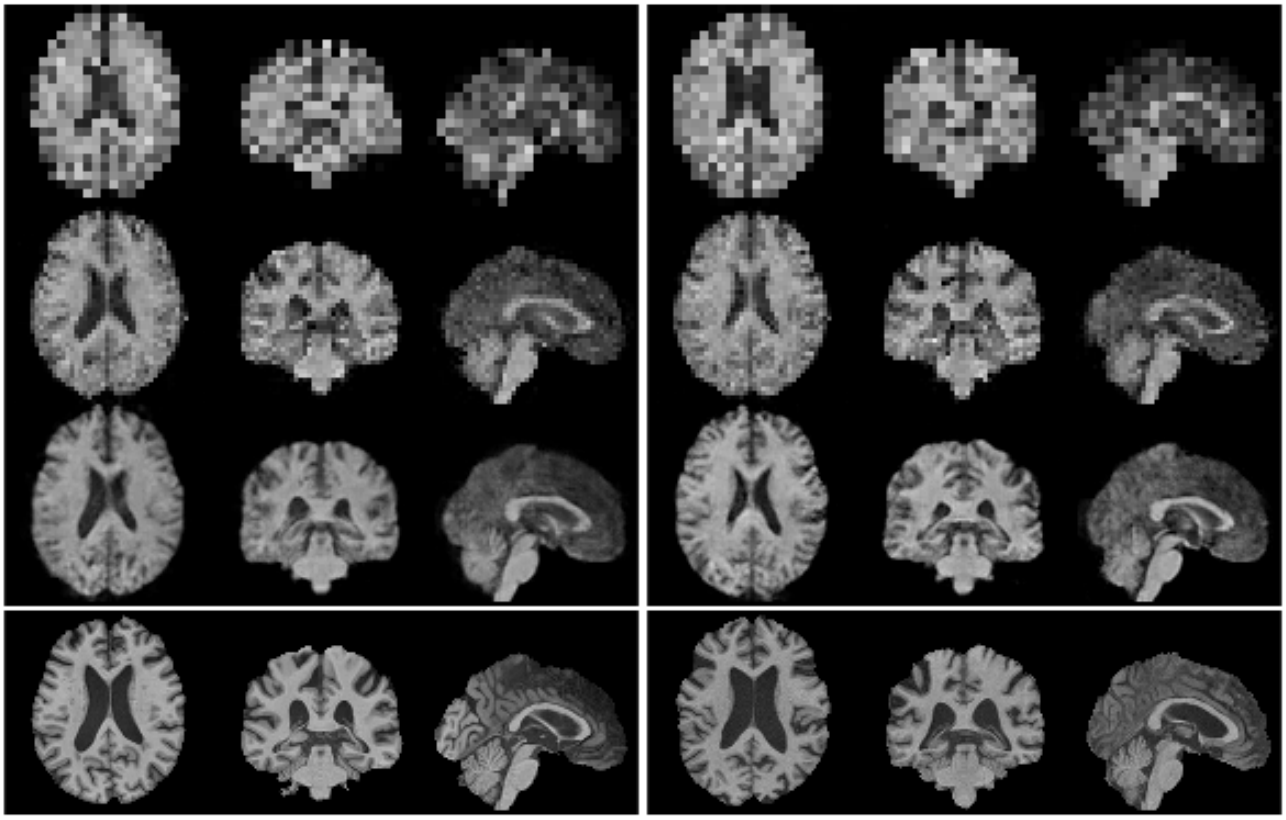


Figure C.7: Results of 3D volume generation. The three stages of increasing image size are shown through three orthogonal slices for two synthetic images, with a pair of sample real images shown (bottom) for comparison. All images re-sampled to the size of the highest resolution generated images (96-by-96-by-96px).

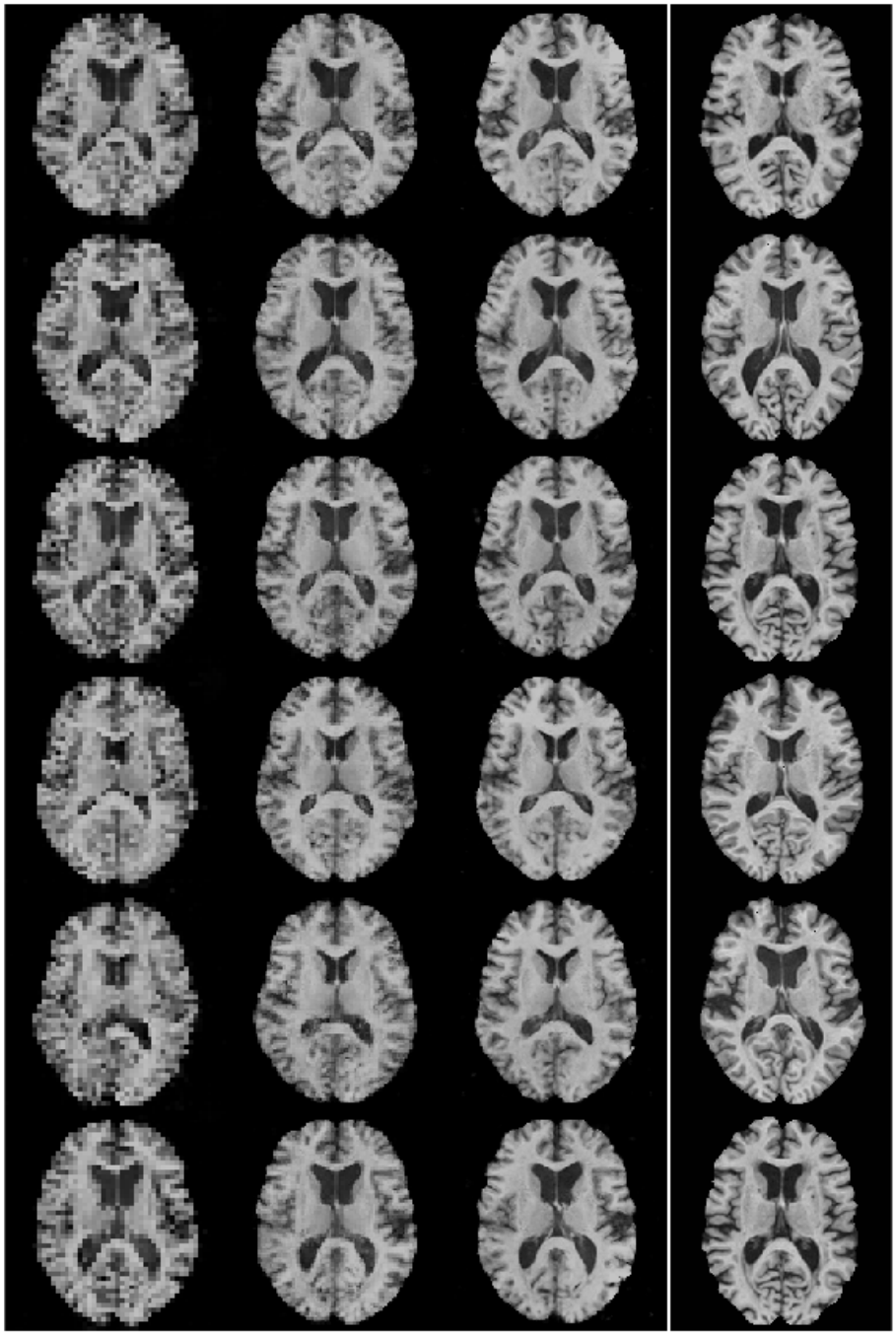


Figure C.8: Results of 2D slice generation. The three stages of increasing image size are shown for 6 synthetic images, with a set of real images shown (rightmost) for comparison. All images re-sampled to the size of the highest resolution generated images (192-by-192px).

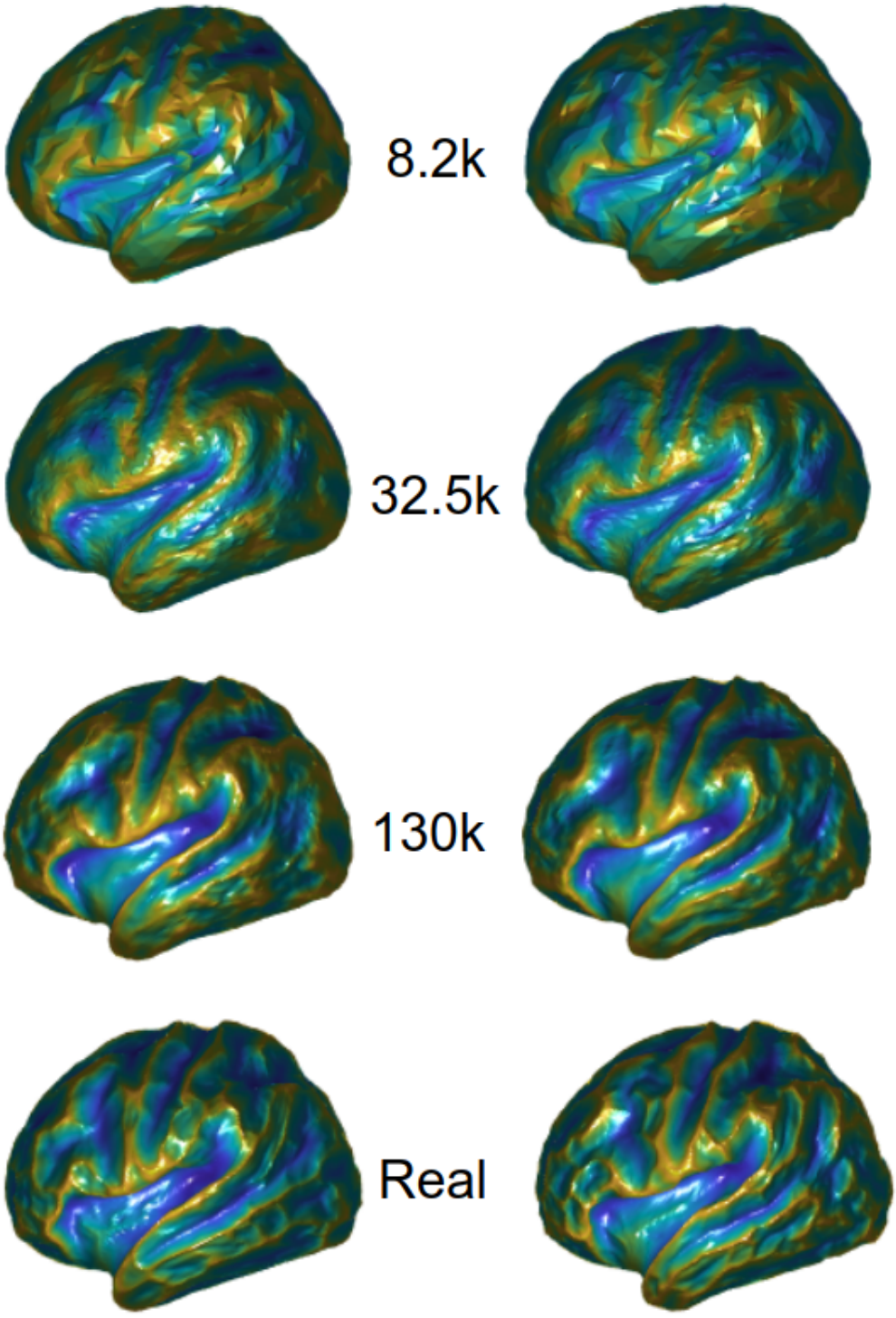


Figure C.9: Results of suclal depth map generation applied to a surface atlas. The three stages of increasing number of vertices (8.2k, 32.5k and 130k) are shown for two synthetic images, with a pair of sample real images shown for comparison.

C.5 Discussion

C.5.1 General appearance

The generated images appear reasonable, with a clear increase in quality from each resolution to the next. Despite this, the highest resolution images are still visibly different from the real images. There appears to be a lack of fine detail, particularly noticeable around the white/grey matter boundaries in the Magnetic Resonance (MR) images. There are no obvious discontinuities between the patches which make up the image, and successively higher resolution images appear to be consistent with the previous lower resolution version. This suggests that local and global consistency have been successfully enforced through equation C.2. However, the lack of high resolution detail is one area in which the proposed method can be improved. This can partially be attributed to the use of an L_1 norm, something which is known to lead to blurry solutions when applied to images, and one of the main reasons why the adversarial loss provided by GANs is so successful. In theory, the realism term in equation C.2 should counteract this effect, however we found that this term had little effect on the images being generated, with there being little difference in the value of this term across the range of generated images. Another factor could be that the GAN does not learn to produce particularly sharp images. An examination of the output of one of the higher resolution GANs in Figure C.10 suggests that this could be the case, with there being a clear distinction between the real and synthetic patches. Since the proposed method is completely independent of GAN formulation, image quality could be improved if more modern or future GAN formulations were used.

C.5.2 Anatomical diversity

Image diversity is an important factor in any generative model. An ideal model should be able to produce images from across the entire manifold of real images. One of the drawbacks to the proposed method is that once the initial low resolution base image is generated, the rest of the process is theoretically deterministic. Whilst the subsequent gradient descent procedures are

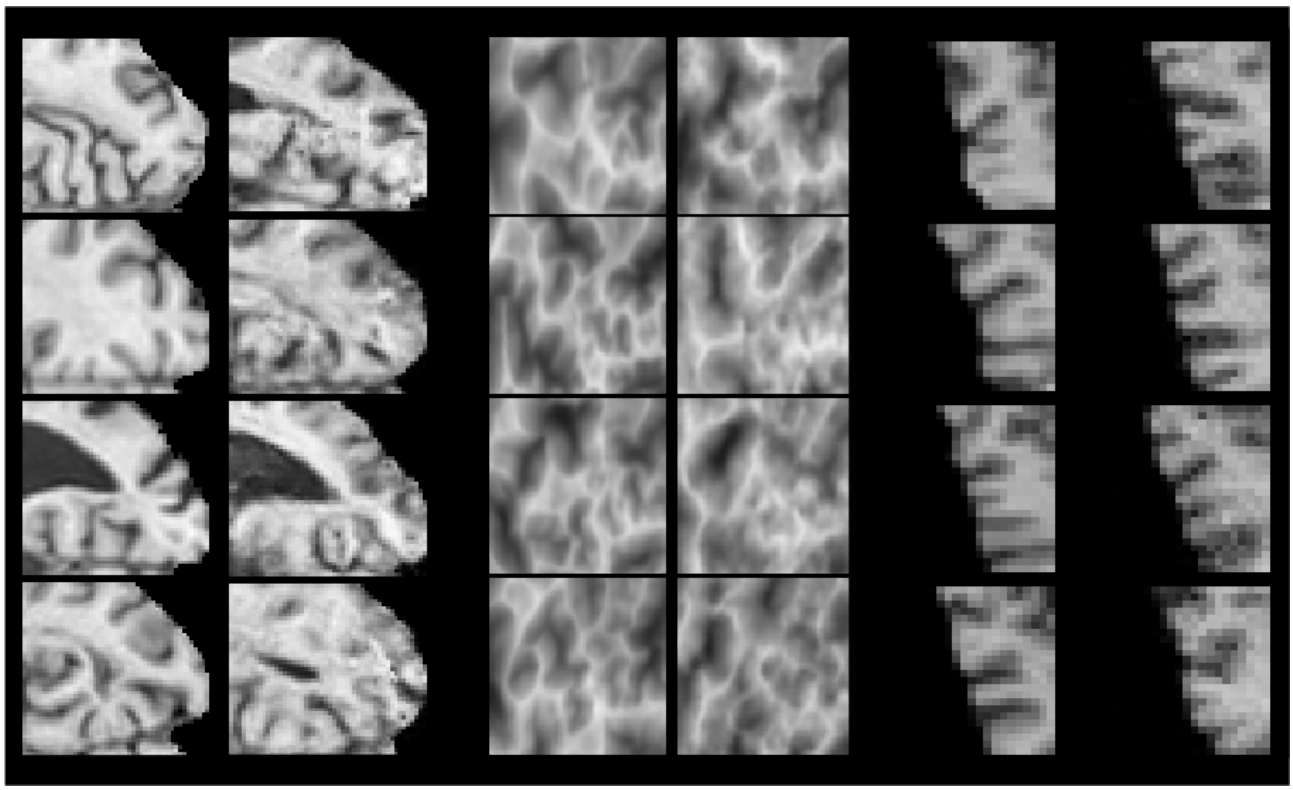


Figure C.10: Example GAN outputs for the highest resolution sub-images for each dataset. Left: 2D MR slice. Middle: Sulcal depth map. Right: Slice through 3D MR volume. In each case the left column shows a random selection of training samples with the right column showing the a random selection of GAN output. Note how image details appear less defined in the synthetic images.

stochastic, the optimal solutions they hope to find are fixed from the moment the base image is generated. This acts to reduce the overall level of diversity since it implies a one-to-one mapping between the low and high resolution image spaces, whereas in reality, there are many valid high resolution images for a given low resolution image. The consequence of this can be seen in Figure C.11 which shows that generated images tend to have a lower variance in ventricle size/shape. One way to develop this method further would therefore be to address this loss of diversity. One possible solution could be to limit the number of components of $\mathbf{z}$ which are allowed to vary during the gradient descent step. Whilst this might lead to higher pixel level errors, it would reintroduce a level of randomness to the generated patches, thereby making the system no longer deterministic.

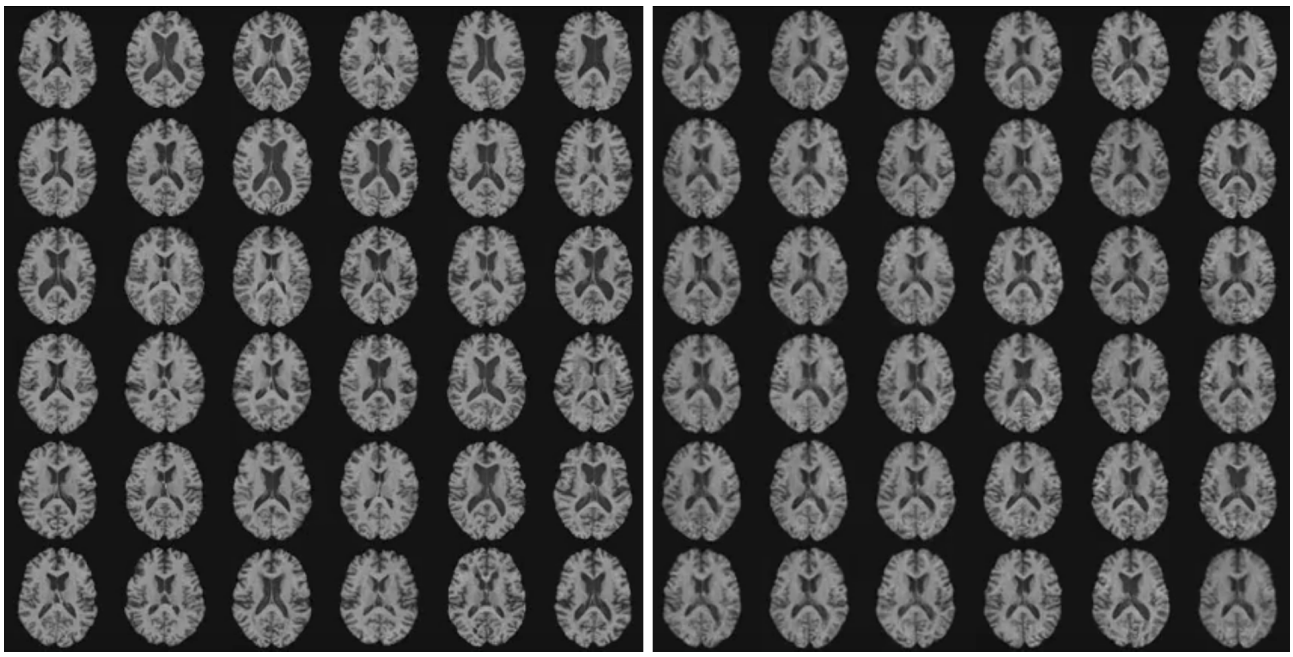


Figure C.11: Left: Axial slices through real images. Right: Axial slices through generated 3D images. Note the lower diversity in ventricle size and shape in the generated images.

C.5.3 Optimisation and run time

Whilst allowing for higher image sizes and surface data to be generated, the proposed method does require long training and induction times. 3D generation involves the training of 27 3D GANs, each taking approximately 20 hours to train. Generating an image then involves 52 rounds of optimisation each taking approximately 30 seconds. Generating a sufficient number of images to calculate any of the objective metrics discussed in Section 3.1.1 is therefore computationally challenging. This, combined with the long training times, means that performing a detailed analysis of the impact of each parameter (GAN architecture and hyper-parameters, patch size, location and overlap, relative influences of the terms in equation C.2 and patch generation hyper-parameters) would be computationally intractable. These parameters were therefore chosen empirically and by using heuristics. This is clearly not ideal and one of the main areas for improvement for the development of the method would be to reduce the run times and number of GANs required. More modern hardware and GAN formulations would both speed up training and inference, and also potentially reduce the number of GANs required by allowing a single GAN cover multiple regions in the high resolution images.

C.5.4 Conclusion

We have presented a flexible method which can be used to allow GANs to generate images beyond simple 2D squares. The method is still in its infancy and would need to be improved in several areas before being applied to a real world problem, however there is potential for the proposed method, or the ideas behind it, to provide a practical way to apply GANs to some areas which would otherwise be inaccessible. One possible area where we see real potential is in the generation of large histology images. These can be extremely large ($>$ 100k-by-100k pixels, see [Alexi et al., 2018] for examples) and are therefore well beyond the range of traditional GANs. Such images also often have high translational invariance, and as such a single GAN could be used at each resolution level, rather than one for each region. This dramatically improves the scalability of the method allowing it to be applied at the extreme image sizes required.

Appendix D

Permissions Table

Material	Source	Copyright holder	Permission requested date	Permission granted	Note
Figure 2.5	Clinical core of the Alzheimer's disease neuroimaging initiative: Progress and plans, *Alzheimer's Dementia: The Journal of the Alzheimer's Association* [Aisen et al., 2010]	Elsevier	Sep 25, 2018	Yes	Permission granted through RightsLink
Figures 3.1 and 3.2	Modality Propagation: Coherent Synthesis of Subject-Specific Scans with Data-Driven Regularization, *Springer eBook* [Ye et al., 2013]	SpringerNature	Sep 25, 2018	Yes	Permission granted through RightsLink
Figure 3.3	Cross-Domain Synthesis of Medical Images Using Efficient Location-Sensitive Deep Network, *Springer eBook* [Van Nguyen et al., 2015]	SpringerNature	Sep 25, 2018	Yes	Permission granted through RightsLink
Figure A.1 and Table A.1	Frontotemporal dementia, *The British Medical Journal* [Warren et al., 2013]	The British Medical Journal	Sep 25, 2018	Yes	Permission granted through RightsLink
Text and figures in Chapter 4	Brain lesion segmentation through image synthesis and outlier detection, *NeuroImage: Clinical* [Bowles et al., 2017]	Elsevier	Sep 25, 2018	Yes	Permission granted through RightsLink
Text and figures in Chapter 7	Modelling the progression of alzheimers disease in mri using generative adversarial networks, *SPIE* [Bowles et al., 2018]	The international society for optics and photonics	Sep 25, 2018	Yes	Shared Copyright

Table D.1: Details of copyrighted work included in this thesis and permissions sought.