Deep Generative Models for Medical Image Synthesis and Strategies to Utilise Them

Tian Xia

A thesis submitted for the degree of Doctor of Philosophy.
The University of Edinburgh.
February 2022

Abstract

Medical imaging has revolutionised the diagnosis and treatments of diseases since the first medical image was taken using X-rays in 1895. As medical imaging became an essential tool in a modern healthcare system, more medical imaging techniques have been invented, such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Computed Tomography (CT), Ultrasound, etc. With the advance of medical imaging techniques, the demand for processing and analysing these complex medical images is increasing rapidly.

Efforts have been put on developing approaches that can automatically analyse medical images. With the recent success of deep learning (DL) in computer vision, researchers have applied and proposed many DL-based methods in the field of medical image analysis. However, one problem with data-driven DL-based methods is the lack of data. Unlike natural images, medical images are more expensive to acquire and label. One way to alleviate the lack of medical data is medical image synthesis.

In this thesis, I first start with pseudo healthy synthesis, which is to create a 'healthy' looking medical image from a pathological one. The synthesised pseudo healthy images can be used for the detection of pathology, segmentation, etc. Several challenges exist with this task. The first challenge is the lack of ground-truth data, as a subject cannot be healthy and diseased at the same time. The second challenge is how to evaluate the generated images. In this thesis, I propose a deep learning method to learn to generate pseudo healthy images with adversarial and cycle consistency losses to overcome the lack of ground-truth data. I also propose several metrics to evaluate the quality of synthetic 'healthy' images. Pseudo healthy synthesis can be viewed as transforming images between discrete domains, e.g. from pathological domain to healthy domain. However, there are some changes in medical data that are continuous, e.g. brain ageing progression.

Brain changes as age increases. With the ageing global population, research on brain ageing has attracted increasing attention. In this thesis, I propose a deep learning method that can simulate such brain ageing progression. Specifically, longitudinal brain data are not easy to

acquire; if some exist, they only cover several years. Thus, the proposed method focuses on learning subject-specific brain ageing progression without training on longitudinal data. As there are other factors, such as neurodegenerative diseases, that can affect brain ageing, the proposed model also considers health status, i.e. the existence of Alzheimer's Disease (AD). Furthermore, to evaluate the quality of synthetic aged images, I define several metrics and conducted a series of experiments.

Suppose we have a pre-trained deep generative model and a downstream tasks model, say a classifier. One question is how to make the best of the generative model to improve the performance of the classifier. In this thesis, I propose a simple procedure that can discover the 'weakness' of the classifier and guide the generator to synthesise counterfactuals (synthetic data) that are hard for the classifier. The proposed procedure constructs an adversarial game between generative factors of the generator and the classifier. We demonstrate the effectiveness of this proposed procedure through a series of experiments. Furthermore, we consider the application of generative models in a *continual learning* context and investigate the usefulness of them to alleviate *spurious correlation*.

This thesis creates new avenues for further research in the area of medical image synthesis and how to utilise the medical generative models, which we believe could be important for future studies in medical image analysis with deep learning.

Declaration of originality

I hereby declare that the research recorded in this thesis and the thesis itself was composed and originated entirely by myself in the University of Edinburgh.

Tian Xia

Acknowledgements

I would like to first thank my supervisor Prof. Sotirios Tsafatris. Thank you Sotos for your valuable guidance throughout these years. Your hard working style and insight into research has inspired me and gave a great example of what makes a great researcher. I will never forget the lessons you taught me and all the good memory we had. I must thank Dr. Javier Escudero and Dr. James Hopgood for their support and useful feedback throughout my PhD. I thank Agis, Chen and Chengjia for inspiring discussions. I thank all my colleagues and friends Andrei, Valerio, Tom, Xiao, Pedro, Greg, Spiros, Marija, Nikolas, Gabriele, Haochuan, Filippo and John for the great time we had in office. I thank all my friedns outside PhD. I cannot thank more my parents for their love and support all these years. Finally, I must give specifical thanks to my other half, my beloved Xi Lin, who has been been by my side throughout the whole PhD, encouraging and supporting me. Without you I would have never been through difficult times.

Contents

List of figures

List of tables

Acronyms and abbreviations

AAE	Adversarial Autoencoder
AD	Alzheimer's Disease
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CJD	Creutzfeldt–Jakob Disease
CN	Cognitively Normal
CNN	Convolutional Neural Network
CSF	CerebroSpinal Fluid
CT	Computed Tomography
DHA	DocosaHexaenoic Acid
DNN	Deep Neural Network
DWI	Diffusion-Weighted Imaging
FLAIR	Fluid Attenuated Inversion Recovery
FID	Fréchet Inception Distance
GAN	Generative Adversarial Network
GBM	GlioBlastoma Multiforme
GM	Gray Matter
GMM	Gaussian Mixture Model
IS	Inception Score
JSD	Jensen-Shannon Divergence
KLD	Kullback-Leibler Divergence
LSGAN	Least Squared Generative Adversarial Network
MAE	Mean Absolute Error

MAP	Maximum-A-Posterior
MCI	Mildly Cognitive Impaired
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NMR	Nuclear Magnetic Resonance
PAD	Predicted Age Differrence
PET	Positron Emission Tomography
PD	Parkinson's Disease
PSNR	Peak Signal to Noise Ratio
RF	Radio Frequency
SSIM	Structural SIMilarity
SVF	Stationary Velocity Field
WGAN	Wasserstein Generative Adversarial Network
WM	White Matter

Chapter 1
Introduction

Medical imaging has revolutionised ways of diagnosing and treating diseases since the first medical image was taken using X-rays in 1895. As medical imaging became an essential tool in the modern healthcare system, more medical imaging techniques have been invented, such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Computed Tomography (CT), Ultrasound, etc. With the advance of medical imaging techniques, the demand for processing and analysing these complex medical images is increasing rapidly. Conventionally, medical images are analysed by clinical experts who have been trained for years, and the processing can be time-consuming depending on the tasks. For example, for medical image segmentation, the medical expert has to label out pixel by pixel where are different organs or where is the target pathology, based on prior medical knowledge. Furthermore, the variations of different imaging techniques and different pathologies, as long as the potential fatigue, introduce more difficulty, risks and expenses into the analysis of medical images. As such, the desire of developing automated computer systems to process medical images has become prevalent.

The core of these computer-assisted techniques is to find informative features that can well represent the patterns inherent in medical images. Previously, these task-related features were mostly designed by human experts based on their own medical knowledge. However, the design of these hand-crafted features can be limited by the variations and biases of these experts, and the designed features may not be suitable for all diseases or imaging techniques. However, the rise of deep learning in recent years has provided an opportunity to overcome this challenge. With more layers and computational power, deep learning models are able to learn these informative features by themselves [17]. As these deep neural networks have the potential to discover inherent features which are difficult for human to notice, they have become the state-of-the-art in many medical imaging analysis tasks, such as pathology segmentation, pathology detection, medical image registration, medical image reconstruction,

etc [18]. However, most deep learning based methods require a large number of medical data and corresponding annotations to train the models. In practice, medical data and annotations are difficult to acquire, which hampers the performance of these DL methods. One way to alleviate the lack of data is to synthesise medical images.

1.1 Motivation

This thesis focuses on one branch of medical imaging analysis, medical image synthesis. In the general computer vision field, the goal of image synthesis is to generate an image that is perceptually realistic. However, in the medical imaging field, in addition to the need for perceptual realism, we also focus on the quantitative accuracy of the synthesised images, i.e. these images have to be clinically meaningful for target tasks such as diagnosis, planning, prognosis, etc. A good definition of medical image synthesis is given as: '*the ability to abstract or summarise (synthesise) knowledge from a collection of examples that are representative of a wider population, phenotype or phenomenon*' [19].

There are many benefits of medical image synthesis. First, medical images are difficult and expensive to acquire. Therefore, deep learning models always have to utilise limited data, e.g. the ISLES dataset [20] only contain 76 volumes while a natural dataset like ImageNet [21] contains millions of images. If we could synthesise medical images with high fidelity, we could perhaps use the synthetic data to improve the training of deep learning models for other tasks (e.g. segmentation, classification, etc.) or even for clinical training purpose. Second, sometimes it is nearly impossible to obtain some medical images, for example, if we have a brain image of a subject who has brain tumour or a neuro-degenerative disease, but we want to see how their brains should look like if they are healthy to evaluate the effect of the diseases. In this case, the 'healthy' version of the diseased brain image is nearly impossible to obtain using CT or MRI machine, because a subject cannot be both healthy and diseased. Similarly, if we want to see how a person's brain should look like in ten years if they have or do not have Alzheimer's Disease, instead of waiting for ten years and then scanning her brain, we could use medical image synthesis to predict their brain images. Furthermore, the ability to synthesise medical images considering different medical factors (e.g. age, gender,

smoking history, or even DNA type) implies that the model has implicitly learnt the complex interplay of these factors and their effect on organs. This could be useful for both research and clinical applications.

As such, the main task of this thesis is to investigate medical image synthesis with different conditions. Specifically, we first investigate 'pseudo healthy' synthesis, i.e. the creation of a 'healthy' image from a pathological one, with the presence or absence of disease as a discrete factor. Then we focus on a more complex task, i.e. the synthesis of aged brain images, with age as a continuous factor and the status of Alzheimer's Disease as a discrete factor. Finally, we show that we can utilise a pre-trained generative model with adversarial training to improve a downstream task..

1.2 Challenges

The first challenge for medical image synthesis is the lack of sufficient training data. Normally, for natural image synthesis tasks, deep learning models are trained on millions of natural images to get good performance. Compared to natural images, medical image data are more difficult to acquire, and the size of medical datasets is always limited. Therefore, it is easy for deep synthesis models to fall into over-fitting or mode collapse due to lack of training data.

The second challenge lies in evaluation. For natural image synthesis such as the synthesis of face images, it is easy for people to tell if a synthetic face image is realistic or not, because people are familiar with and sensitive to faces. However, for medical image synthesis, as most people are not familiar with MRI or CT images, it requires either medical experts or ground-truth target images to evaluate the synthetic results. Inviting medical experts to help is not easy and is expensive, and in some cases, there are no ground-truth target images, e.g. the synthesis of a 'healthy' image from a pathological one.

The third challenge of medical image synthesis are the strict requirements for the synthetic images. For natural image synthesis, most methods only need to show the fidelity of the synthetic images. However, for medical image synthesis, the synthetic images have to be

both realistic and clinically meaningful.

1.3 Overview and Technical Contributions

An overview of the contributions of this thesis is provided below. Recall that the ultimate goal of this work is to find methods that can synthesise medical images under different clinical conditions. We first start with a simple task, i.e. synthesising a 'healthy' brain image from a pathological one. Then we proceed by involving 'age' as a continuous factor and simulate 'aged' brain images given different target ages. At last, we propose an adversarial procedure to utilise pre-trained generative models for downstream tasks.

In chapter 4, an adversarial deep learning model is proposed to synthesise 'healthy' images from a pathological image. The proposed approach mainly contains three components: the Generator, the Segmentor and the Reconstructor. The Generator transforms a pathological image to a 'healthy' image, and adversarial loss is used to train the Generator as there is no ground-truth available. To maintain subject identity, the cycle-consistency loss is applied. However, when transforming between a domain that contain more information (e.g. pathological domain) and a domain that has less information (e.g. healthy domain), there is a problem called 'one-to-many' problem. Because if the 'healthy' image is truly healthy, then there will be not diseased information in the 'healthy' image to reconstruct the diseased image. Normally, the deep model tends to hide the pathological information in 'healthy' image to allow reconstruction, which affects the image quality. To alleviate the 'one-to-many' problem, the Segmentor is used to obtain the pathological information in the form of binary segmentation. Then the Reconstructor combines the 'healthy' image and pathological segmentation to reconstruct the pathological image. The contribution of this chapter is a new method to synthesise 'healthy' image by solving the *one-to-many* problem.

Chapter 4 is based on the following publications:

- Tian Xia, Agisilaos Chartsias, and Sotirios A. Tsaftaris. "Adversarial Pseudo Healthy Synthesis Needs Pathology Factorization". In International Conference on Medical Imaging with Deep Learning, pp. 512-526. PMLR, 2019.

- Tian Xia, Agisilaos Chartsias, Sotirios A. Tsaftaris, "Pseudo-healthy synthesis with pathology disentanglement and adversarial learning", Medical Image Analysis, Volume 64, 2020, 101719, ISSN 1361-8415.

The code for this chapter is publicly available at `https://github.com/xiat0616/pseudo-healthy-synthesis`.

In Chapter 5, a deep learning model is proposed to simulate the ageing of the brain without longitudinal data. The model mainly contains an Encoder, a Decoder and a Transformer. The Encoder extracts the anatomical features of a young brain images, and the Transformer makes changes to the features conditioned on a target age, and then the Decoder produces the aged images. The contribution of this chapter is the first model to achieve brain ageing synthesis without the use of longitudinal data.

Chapter 5 is based on the following publications:

- Tian Xia, Agisilaos Chartsias, and Sotirios A. Tsaftaris, for the Alzheimer's Disease Neuroimaging Initiative. "Consistent Brain Ageing Synthesis". In: Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11767. Springer, Cham. https://doi.org/10.1007/978-3-030-32251-9-82
- Tian Xia, Agisilaos Chartsias, Chengjia Wang, Sotirios A. Tsaftaris. "Learning to synthesise the ageing brain without longitudinal data", Medical Image Analysis, Volume 73, 2021, 102169, ISSN 1361-8415. .

The code for this chapter is publicly available at `https://github.com/xiat0616/BrainAgeing`.

In Chapter 6, a simple approach is proposed to utilise pre-trained generative models for downstream tasks. Specifically, I choose the classification of Alzheimer's Disease as the downstream task and use the brain ageing model in Chapter 5 to improve the classification. The proposed approach formulates an adversarial game between the *conditional factor* (target age), on which the generative model is conditioned, and the *AD classifier*. This can be

viewed as finding the 'weakness' of the AD classifier and force the classifier to overcome its 'weakness'. The contribution is a simple adversarial approach to utilise pre-trained generative models.

Chapter 6 is to be submitted to MICCAI 2022 or a CV conference.

1.4 Thesis structure

An overview of the thesis structure is provided below. Chapter 2 introduces medical background. Chapter 3 summarises the technical background. Chapter 4 describes the work on 'pseudo healthy' synthesis. Chapter 5 describes the work of brain ageing synthesis. Chapter 6 introduces the adversarial procedure to utilise pre-trained generative models. Finally, Chapter 7 concludes the manuscript, discussing limitations and future works.

Chapter 2
Medical Imaging and Clinical background

This thesis mainly focuses on Magnetic Resonance Imaging (MRI), a non-invasive technique using magnetisation to image soft tissues. Although the machine learning approaches in this thesis do not consider the physics of the MRI image acquisition, some fundamentals of MRI are briefly introduced in Section 2.1. Since this thesis mainly focuses on brain image analysis, a general introduction of brain anatomy is provided in Section 2.2. Then a background of several brain pathologies is presented in Section 2.3, as well as a background of brain ageing in Section 2.4. Finally, a summary of the brain MRI datasets used in this thesis is presented in Section 2.5.

2.1 Magnetic resonance imaging

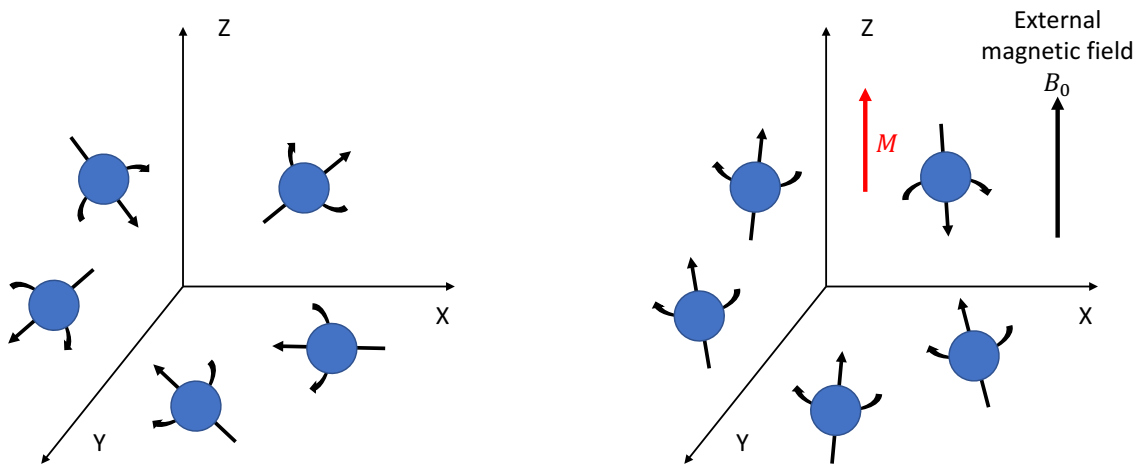
MRI has played an increasingly important role in modern healthcare due to its noninvasive characteristic and ability to generate versatile contrasts for different organs. The history of MRI can be dated back to 1946, when Bloch and Purcell independently discovered the nuclear magnetic resonance (NMR) phenomenon, for which they were awarded the Nobel Prize for Physics in 1952 [22, 23]. In the 1970s, Lauterbur et al. [24] and Mansfield et al. [25] made fundamental contributions to MRI making its clinical applications a reality, for which they were awarded the Nobel Prize for Medicine in 2003. As techniques evolved rapidly, the first clinical MRI images were produced in Nottingham and Aberdeen in 1980 [26, 27]. Since then, MRI has become an essential clinical tool.

The secret of MRI lies in hydrogen, which is the most abundant element in the human body. The hydrogen nucleus, ^1H, possess a property known as “spin”, which can be analogously conceived as the nucleus spinning around its own axis. The spin of the hydrogen nucleus

generates a local magnetic field. Without an external magnetic field, the directions of these local magnetic fields are random, as shown in Figure 2.1(a). However, when a strong, external magnetic field B_0 is applied, the protons will be aligned either in parallel with or anti-parallel to the external field, as shown in Figure 2.1(b). This results in a macroscopic magnetization M parallel to the external field. The protons spin in one of two energy states: a low-energy state (oriented parallel to the magnetic field) and a high-energy state (orientated anti-parallel to the magnetic field direction). Note that the protons do not orient completely parallel or anti-parallel to the field but rotate around the field direction at a frequency known as "Larmor frequency". The Larmor frequency is given by:

$$f_L = \frac{\gamma}{2\pi} B_0, \tag{2.1}$$

where γ is the gyromagnetic ratio which is a constant for a specific nucleus, *e.g.* for protons, $\gamma = 267.5$ MHz/T. When nuclei are placed under the static magnetic field B_0, they can be excited by the application of an electromagnetic radiofrequency pulse (RF pulse) B_1 which is applied perpendicular to B_0 and oscillates at the Larmor frequency. As shown in Fig-



(**a**) No external magnetic field.

(**b**) External magnetic field in Z direction.

Figure 2.1: Illustration of protons' magnetic moment. (a) When there is no external magnetic field, the protons' magnetic moments are in random directions; (b) when an external magnetic field is applied, the protons' magnetic moments precess around the axis of the magnetic field.

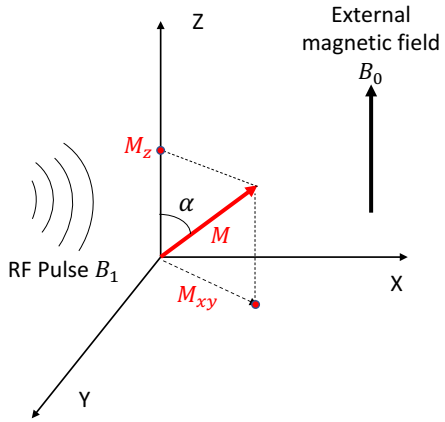


Figure 2.2: The RF pulse B_1 tilts the macroscopic magnetisation M towards the xy plane with a flip angle α, resulting in a non-zero component M_{xy} in the xy plane.

ure 2.2, the oscillating magnetic field B_1 will tilt the macroscopic magnetisation M towards the transverse xy plane with a flip angle α that depends on the duration and amplitude of B_1. Consequently, M will have a component in the xy plane, *i.e.* M_{xy}, and a component in the z axis, *i.e.* M_z. This process is called the excitation phase, during which a number of nuclei absorb energy and change from low energy state to high energy state.

After the RF pulse B_1 ends, the nuclei return to the initial equilibrium state, with the macroscopic magnetisation M returning to its original state. This process is called the relaxation phase. Specifically, M_z regrows to its equilibrium value M, and M_{xy} decays to its original value of zero, which can be described by:

$$M_z(t) = M_z(0)(1 - e^{-\frac{t}{T_1}})$$
$$M_{xy}(t) = M_{xy}(0)e^{-\frac{t}{T_2}},$$

where T_1 is called the T1 relaxation time, or *spin-lattice relaxation time*, after which about 63% ($1 - 1/e \approx 63\%$) of the magnetisation alongside z axis has recovered; and T_2 is called the T2 relaxation time, or *spin-spin relaxation time*, after which the transverse magnetisation decreases to about 37% of $M_{xy}(0)$. During the relaxation process, nuclei release the absorbed energy in electromagnetic waves that can be measured by the MRI scanner using multiple RF receiver coils. The MRI scanner does not acquire the image directly but gathers the information of the signal in frequency or K-space. Finally, the medical image is obtained

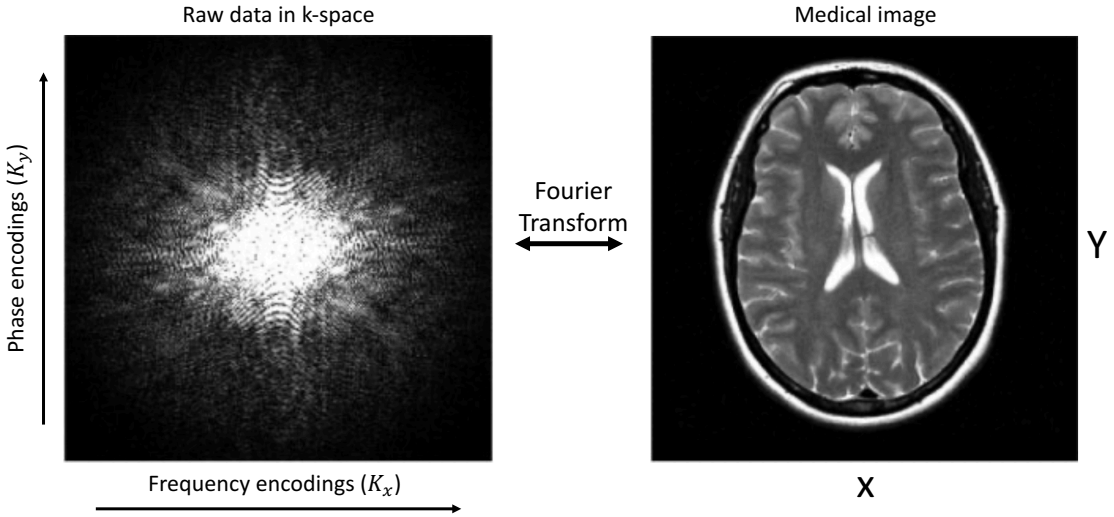


Figure 2.3: Example of a brain MRI image obtained from K-space. Image is taken from [1].

from the K-space array by inverse Fourier transform. An example K-space array and its corresponding medical image are shown in Figure 2.3. A K-space array is a 2D array that is comprised of vectors (or points) (k_x, k_y) containing spatial frequency information of the image.

To determine the spatial location of the received signal, MRI requires the application of additional magnetic field gradients, which leads to spatially varying magnetic fields. Different RF pulses and magnetic gradients form different MR sequences, resulting in different image contrasts. Commonly used MR sequences include T1-weighted, T2-weighted, Fluid Attenuated Inversion Recovery (FLAIR) weighted and the Diffusion-weighted sequences, etc. With this variability, an MRI scanner can obtain medical images with various contrasts (or modalities) for different clinical and research purposes. However, MRI images are expensive and time-consuming to obtain, resulting in the limited size of available MRI datasets that can be used to train machine learning methods. Therefore, synthesising MRI data that are realistic and diverse can be helpful for MRI analysis with machine learning. This thesis uses brain MRI images as research materials to train and evaluate our deep learning methods. In the following section, we introduce some basic knowledge about the brain, the organ of focus in this thesis.

2.2 Brain anatomy

The human brain can be divided into three parts: *cerebrum*, *cerebellum* and *brainstem* [28]. An example image of the brain with marked *grey matter*, *white matter* and *lateral ventricles* is shown in Figure 2.4. The cerebrum is the largest part of the brain and consists of two parts, *i.e.* the left and right cerebral hemispheres. The surface of each cerebral hemisphere is called the cerebral cortex, which consists of billions of neurons and is central to cognitive activities, *e.g.* motor function, language processing, determining personality, etc. The cerebral cortex is also called *grey matter* due to its greyish brown appearance. The cortex has a folded appearance that allows more neurons to fit inside the skull. Each fold is called a *gyrus*, and each valley between folds is a *sculcus*. Beneath the cortex are long nerve fibres connecting brain areas to each other, which affects learning and is called *white matter* due to its relatively light appearance. There are hollow cavities inside the brain, called *ventricles*, which are filled with *cerebrospinal fluid* (CSF), a fluid that flows within and around the brain and protects it.

The cerebellum is also called *little brain* and is located under the cerebrum [29]. The functions of the cerebellum include maintaining balance, coordinating muscle movements, and working memory. Like the cerebrum, the cerebellum is made up of two hemispheres, which are connected through *vermis*. The brainstem is located in front of the cerebellum and is made up of three structures: *pons*, *medulla oblongata* and *midbrain* [30]. It plays a role as a relay station connecting the cerebrum and cerebellum to the spinal cord. The brainstem is the centre of many primitive functions that are essential for survival, such as breathing, heart rhythms, swallowing, etc.

2.3 Brain pathology

In Chapter 4, we propose a deep learning method to perform pseudo healthy synthesis, *i.e.* generating subject-specific 'healthy' images from pathological ones. To train and evaluate the proposed method, we use two brain MRI datasets covering ischemic stroke and brain tumours. Here we provide a brief background of stroke in Section 2.3.1 and brain tumours in Section 2.3.2.

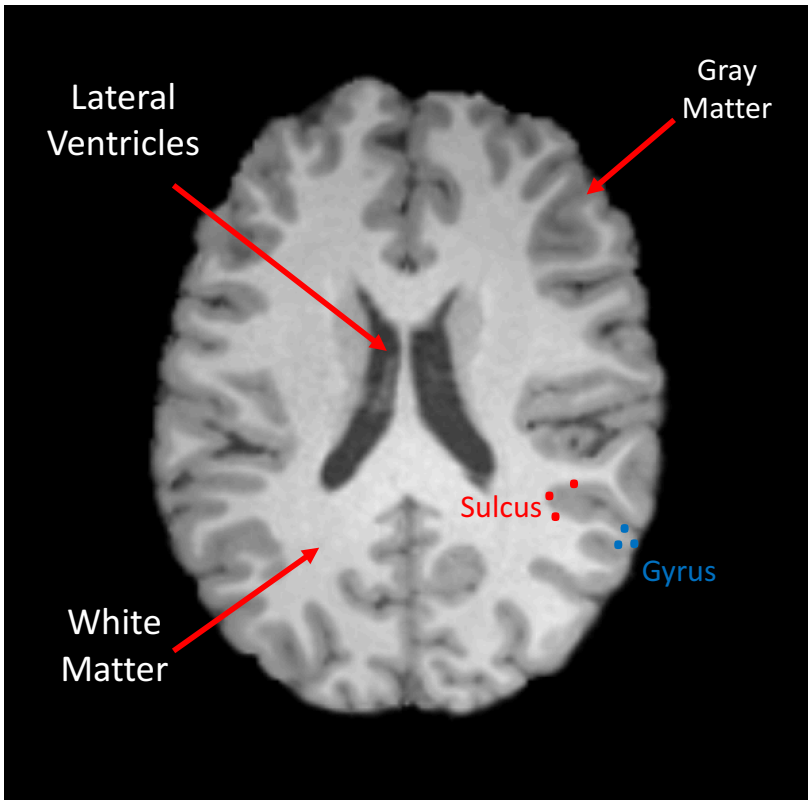
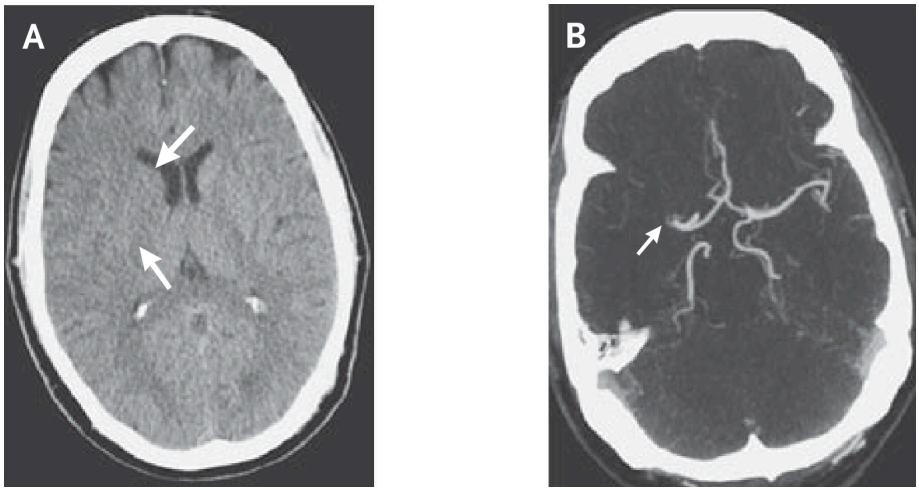


Figure 2.4: Example of a brain T1-weighted MRI image. Lateral ventricles, grey matter and white matter are marked with red arrows; examples of *sulcus* and *gyrus* are dotted in red and blue, respectively. This image is taken from the Cam-CAN dataset [2].

2.3.1 Stroke

Stroke is a leading cause of death worldwide, ranking second after ischemic heart disease [31]. A stroke happens when the blood supply to part of the brain is interrupted or reduced, hindering brain tissues from getting oxygen and nutrition and resulting in brain cell deaths in hours or even minutes. There are three major types of strokes differing in their causes. *Ischemic stroke* is the most common stroke, accounting for about 85% of all cases, and is caused by the blockage in the blood vessels of the brain. The second most prevalent stroke is *Haemorrhagic stroke*, caused by the blood vessel burst of the brain. Another type of stroke, *Transient ischemic attack* (sometimes called 'mini stroke'), is similar to ischemic stroke, and the difference is that the brain is only blocked for a short period, normally for several minutes. As this thesis uses brain data of ischemic stroke, we mainly focus on this stroke in the rest of this



(**a**) Brain CT image with ischemic stroke

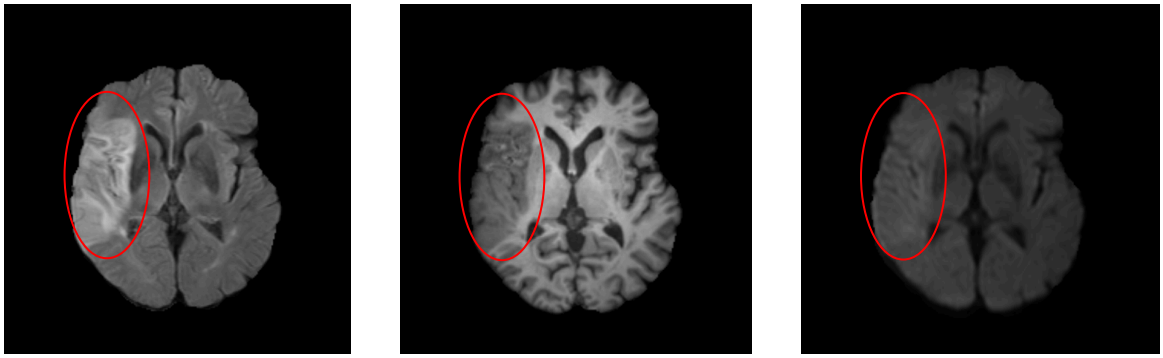
(**b**) Brain CT angiographic image with ischemic stroke

Figure 2.5: Example of a brain with ischemic stroke: (a) a brain CT image with arrows pointing out slight abnormal differentiation of grey and white matter in the basal ganglia; (b) a brain CT angiographic image with arrows showing the occlusion of the first segment of the right middle cerebral artery. Images are taken from [3].

section.

Stroke can occur at any age, although the incidence of stroke and a poor outcome increases markedly with age [32]. Age is *de facto* one of the most important risk factors for all types of stroke, including ischemic stroke [33]. It has been reported that 75-89% of strokes occur after the age of 65 [34], and the likelihood of stroke doubles every successive decade after 55 years old [35]. As the global population is ageing, the worldwide impact and cost of stroke are increasing as well, highlighting the importance of studies on the treatment and prevention of stroke.

Typical signs and symptoms of ischemic stroke include vision problems, sudden numbness or weakness in face, arm or leg, sudden confusion or loss of coordination, etc. Ischemic stroke is typically characterised and initially assessed by sudden onset of neurological impairments. However, diagnosis based on signs and symptoms is not certain, and computed tomography (CT) and MRI are required for further diagnosis and treatment. CT is faster, more widely available and less expensive than MRI, but MRI is more sensitive for detecting ischemic stroke, especially in the first hours after it occurs [35]. Figure 2.5 presents an example of



(**a**) FLAIR (**b**) T1 (**c**) DWI

Figure 2.6: Example MRI images of a brain with ischemic stroke in (a) FLAIR, (b) T1 and (c) DWI modalities. We can observe the stroke in all modalities, marked by red circles. Images are taken from Ischemic Stroke Lesion Segmentation (ISLES) challenge 2015 dataset.

brain CT and CT angiographic images with ischemic stroke, where we can observe slight signs of stroke. Figure 2.6 shows a brain with ischemic stroke in FLAIR, T1 and DWI - weighted MRI images, from which we can clearly observe the shape and location of the brain region affected by the ischemic stroke.

Successful treatment of ischemic stroke requires fast diagnosis involving the acquisition of the information of the stroke lesion presence, extent, location and other factors from brain medical images. Therefore, an automated approach that can locate, segment and quantify the brain lesion areas is of great value, which is the motivation of the Ischemic Stroke Lesion Segmentation (ISLES) challenge 2015. However, such automated approaches often suffer from limited training data, a problem that besets the medical imaging research community, which could be alleviated by medical image synthesis by generating realistic data. In Chapter 4, we propose a method that generates 'healthy' images from pathological ones, and we validate our method on a brain MRI dataset that contains patients with ischemic stroke.

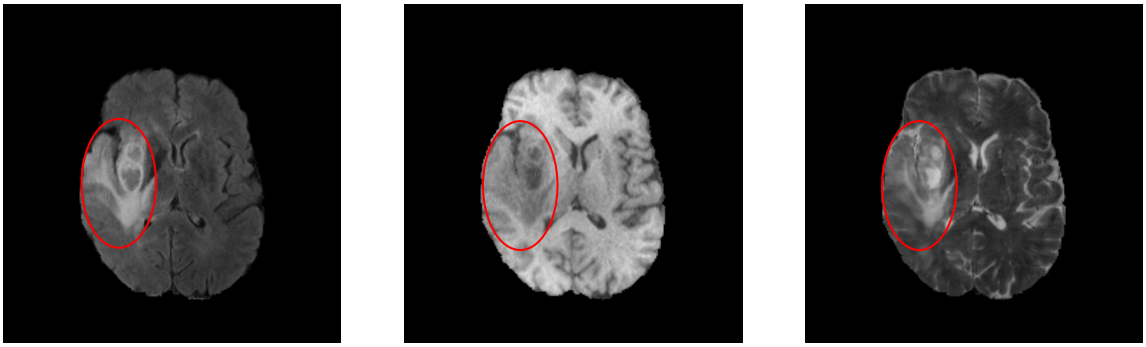
2.3.2 Brain tumour

In Chapter 4, we also validate our method on brain images with tumours. Here we give a brief introduction to brain tumours. Brain tumours result from abnormal growth of cells within the brain or supporting tissues and can heavily impair the brain and its function, posing a severe

threat to life. Brain tumours can be categorised into four grades based on their severity: grades 1 and 2 are low-grade or *benign*, while grades 3 and 4 are high-grade or *malignant*. The main difference between benign and malignant brain tumours is that malignant brain tumours spread quickly in the brain and seriously threaten life, while benign brain tumours are milder with a slower growth rate and do not pose a serious threat to life at least at the time of assessment.

There are more than 150 different types of brain tumours that can be mainly classified into two groups: *metastatic* and *primary*. Metastatic brain tumours originate from somewhere else in the body and then transfer to the brain, typically via the blood vessels. This type of brain tumour is malignant and recognised as cancer. In fact, almost one-quarter of cancer patients suffer from metastatic tumours to the brain. By contrast, primary brain tumours originally occur in the brain or its supporting tissues. The primary brain tumours can be further grouped as glial (if consisting of glial cells) and non-glial (if developed in the structures of the brain, *e.g.* nerves and blood vessels). Unlike metastatic brain tumours, a primary brain tumour can be malignant or benign, depending on its extent and impact.

Gliomas are the most common primary brain tumours, accounting for more than three-quarters of malignant brain tumours. These tumours develop from a kind of supporting cells called *glia*, which can be subdivided into *astrocytes*, *ependymal cells* and *oligodendroglial cells* (or *oligos*). Depending on which glial cells they originate from, gliomas can be further categorised into different subtypes. Among them, *astrocytomas* are the most frequent gliomas, accounting for nearly 50% of primary brain tumours. This category of gliomas grows from astrocytes, a kind of star-shaped glial cells, which are part of the supportive tissue of the brain. Astrocytomas mostly occur in the cerebrum and can develop at all ages but more in adults. For children, these tumours are mostly low-grade, while for adults most are high-grade. *Eponymous* are much less frequent than astrocytomas, accounting for only two to three per cent of brain tumours. These tumours develop from ependymal cells that line the ventricular system, and most of them are benign. *Glioblastoma multiforme* (GBM) is the most aggressive type of glial tumours, which normally grows and spreads to other tissue rapidly, resulting in poor outcomes. This kind of gliomas may consist of different types of cells, such as astrocytes and oligodendrocytes. Other types of gliomas include *oligoden-*



(**a**) FLAIR (**b**) T1 (**c**) T2

Figure 2.7: Example MRI images of a brain with a tumour in (a) FLAIR, (b) T1 and (c) T2 modalities. We can observe the tumour in all modalities, marked by red circles. Images are taken from Multimodal Brain tumour Segmentation (BraTS) 2018 Challenge [4].

drogliomas which are derived from the cells comprising myelin and *Medulloblastomas* which usually develop in the cerebellum.

Common signs and symptoms of brain tumours include headache, vomiting, drowsiness, mental decline, behaviour changes, vision or speech problems, etc. Although the exact cause of brain tumours is still unknown, there are some identified risk factors. The incidence of getting a brain tumour increases with age, while some brain tumour types are more common in children [36]. Exposure to therapeutic doses of ionizing radiation is also found associated with brain tumour risk [37]. Moreover, some genetic conditions are found to increase the risk [36]. The diagnosis of brain tumours may involve inquiry of the above-mentioned symptoms and risk factors, but the definite diagnosis will need the use of medical imaging techniques, especially MRI. Although CT is cheaper and faster than MRI, it can neglect some structural lesions, particularly in the posterior fossa, or non-enhancing tumours like low-grade gliomas [36]. By contrast, MRI can provide more image details and thus becomes the better choice for brain tumour diagnosis despite its higher expense and less availability. Examples of MRI images of brains affected by tumours are presented in Figure 2.7. Note that the tumours may displace normal brain tissues. The deformation induced by growing brain tumours is so-called *mass effect*, which distinguishes brains affected by tumours from those affected by stroke. In Chapter 4, we propose a deep learning-based procedure that aims to generate 'healthy' images with fixing deformations induced by brain tumours.

2.4 Brain ageing

In Chapter 5, we propose a deep learning-based model that can simulate brain progression by synthesising MRI brain images. Here we briefly introduce brain ageing, as well as the anatomical and cognitive changes of this process.

An ageing population across the world has brought several challenges and burdens to modern healthcare systems [38]. Although the ageing process is complex and its underlying mechanisms are still being investigated, it is suggested that ageing increases the risks of major human pathologies, including cancers, cardiovascular disorders and neurodegenerative diseases [39]. Among all the negative aspects of growing old, deterioration of brain function is probably the most fearful one for many people, especially those approaching older. The human brain constantly changes structurally and functionally throughout the whole lifespan. Even for elderly people considered cognitively normal, their brains still go through age-related changes resulting in brain volume reduction and function degeneration.

2.4.1 Cognitive decline

Brain cognitive abilities can be classified into two types: *crystallised abilities* and *fluid abilities* [40, 41]. Crystallized abilities refer to the skills and knowledge accumulated from past learning and experience [41]. Typical examples of crystallised abilities include verbal ability, mathematical skills and general knowledge. Fluid abilities refer to the ability to think, reason and solve problems, with minimal dependence on previous education, skills and experience [41]. Typical examples of fluid abilities are classifying figures, solving puzzles and coming up with problem-solving strategies. In short, fluid abilities determines how fast and well we learn new knowledge, and the learnt knowledge forms the crystallised abilities.

Ageing has different effects on these two kinds of cognitive abilities [42]. Specifically, the fluid abilities such as processing speed, reasoning and memories decline from young age onward [43]. In contrast, the crystallised abilities are less affected by ageing until very late age [44, 45]. As a consequence, people find it harder to memorise new knowledge and process new tasks since early adulthood, but once they acquire new skills and new knowledge, their

crystallised intelligence increases [46]. However, after a specific age, e.g. age 60, they begin to forget things that they have learnt in early life, and crystallised mental abilities decrease [44].

Among all age-associated cognitive changes, memory deterioration might be the most observed one, particularly for elderly people [47, 48]. Specifically, *episodic memory* and *semantic memory* are the two sections of memory function that are most affected by ageing [48]. The episodic memory is the mental capacity to recall and re-experience episodes of one's personal life, *i.e.* when, where and how things happened [49, 50]. An example of episodic memory could be that on the first day of primary school, you learnt bread is made from flour. Semantic memory is a long-term memory of concept-based knowledge [49, 50]. An example of semantic memory could be knowing that bread is made from flour. It has been found that both episodic and semantic memories decline from middle age onward. Other typical age-associated cognitive changes include slower reaction times, lower attention levels, slower mental speed, worse perceptual functions, etc [48].

However, why and how these cognitive changes occur is not fully explored yet [48, 51, 52]. The development of MRI has offered opportunities to study the relationship of cognitive decline and structural changes of the brain during ageing [47, 53, 54].

2.4.2 Structural changes

Studies have shown that there is a decline in brain volume and weight as age grows at a gross level, accompanied by an increase in ventricular volume and cerebrospinal fluid [55]. However, ageing does not affect the brain uniformly. The ageing trajectories for different brain structures are different. Some brain structures are preserved well even in late life, while others decrease substantially [47, 56]. Some brain structures start to change even from early adulthood, while others deteriorate after mid age [47, 56].

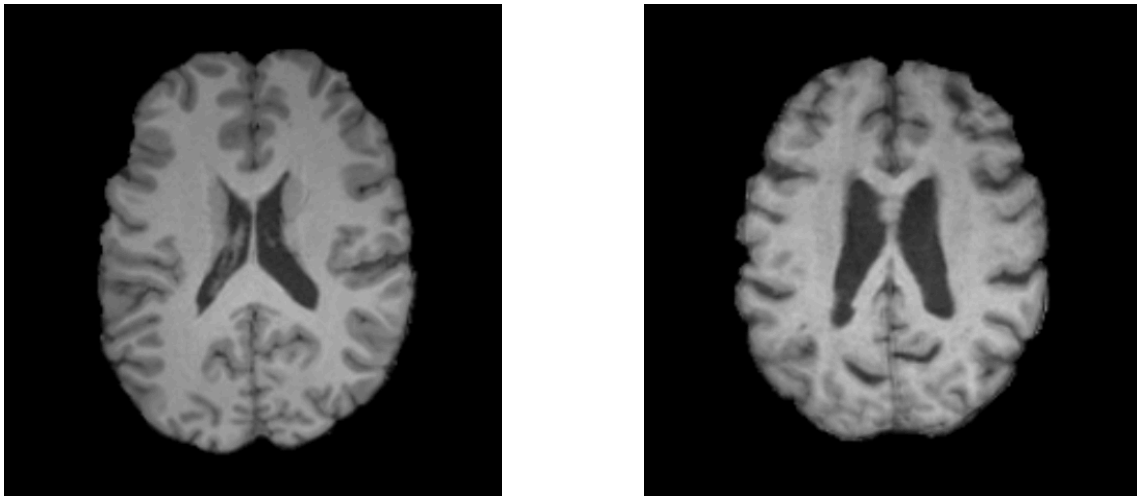
The grey matter (GM) has been found to decrease with age even, and the decline begins in early life [57, 47, 58, 59]. Specifically, the prefrontal cortex suffers from the most prominent atrophy, followed by frontal cortex [47, 59]. This is consistent with previous studies that executive functions, which heavily depend on the frontal neural circuits, are the most affected

by ageing among cognitive functions [47, 53]. The temporal cortex and other GM areas are also affected by ageing though the effect is less than that of the frontal cortex [47, 60].

White matter (WM) plays an important role in central cognitive activities, as it transfers information between different cortical areas [47, 61, 62, 63, 64]. The deterioration of WM could impair the integration of information from distant cortical areas. It has been found that the effect on WM is different from on GMr [65, 66, 67, 68]. The WM volume remains relatively stable since adulthood but starts to decrease after about age 70 [69, 70, 71]. Although the reduction of WM volume occurs later than that of GM, the WM atrophy is more rapid than GM atrophy and accelerates with age [72, 73]. Eventually, the WM atrophy exceeds the GM atrophy [47, 73].

The cerebellum is less affected by ageing than the cerebrum [74, 75, 47]. To be specific, the cerebellar WM observes a linear development for the first part of the lifespan, and the decline accelerates after a relatively old age [76, 47]. The cerebellar GM, on the other hand, is found linearly correlated with age, though the decrease rate is relatively slow [47]. Studies also found that hippocampus is affected by aging, possibly resulting in reductions in episodic memory [77, 78, 79, 80].

Examples of a young brain and an old brain are shown in Figure 2.8. It is now widely accepted that brain volume reduces as age grows, but such reduction is not uniform. Brain structures have different ageing trajectories. Studies have shown that structural changes of some brain regions are correlated with specific cognitive performances [78, 47, 81]. However, more efforts are needed to better reveal the relationship between brain structural changes and cognitive decline. Most previous works were cross-sectional studies that suffer from confounding caused by cohort difference [82, 83, 84]. Different generations could differ in their lifestyles, culture, education and daily diet, implicitly affecting the cross-sectional conclusions [47]. Even the longitudinal studies are subject to cohort bias [85]. The subjects that remain in the follow-up study tend to be the healthiest and have the best cognitive scores [85]. Besides, since the subjects repeat the same test more than once, it is possible that they could maintain or even improve the test scores even with cognitive decline [85, 44]. The limitations of these studies are the motivation of our work in Chapter 5, where a deep learning method is proposed to simulate subject-specific ageing trajectories age by age.



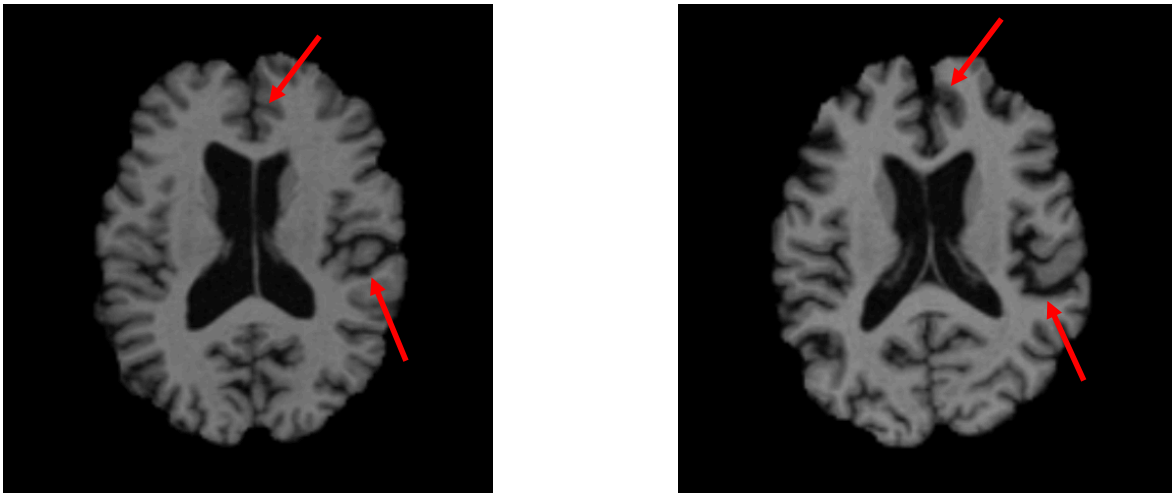
(**a**) A brain at age 24 (**b**) A brain at age 80

Figure 2.8: Examples of a young brain and an old brain. Structural changes such as volume reduction and ventricular enlargement can be observed. Images are taken from CamCAN [2].

2.4.3 Neurodegenerative diseases

Apart from normal ageing, the brain could suffer from neurodegenerative diseases that cause pathologically structural and cognitive changes. Typical examples of neurodegenerative diseases include Alzheimer's disease (AD), Parkinson's disease (PD), Huntington's disease, Creutzfeldt–Jakob disease, etc. Among them, the most prevalent neurodegenerative disease is AD, followed by PD.

Similar to normal ageing, AD is associated with a series of cognitive and structural changes. Since normal ageing affects most of the brain, it is not realistic to find brain areas that are affected only by AD but not by normal ageing. In fact, age is an important risk factor for AD. The brain changes caused by AD are normally intertwined with those by normal ageing, with similarities and differences. In general, the brain regions that are affected by AD undergo accelerated atrophy compared to normal ageing, but the effect is not uniform: some brain regions are more affected by AD and degenerate faster than others. Specifically, the effects of AD are most predominant in the medial temporal structures (*e.g.* the hippocampus, entorhinal cortex, retrosplenial cortex, parahippocampal gyrus, etc.) that play a crucial role in episodic memory [47, 86]. This is in line with previous studies suggesting that AD heavily impacts the episodic memory [87, 88]. As a result, a typical event for AD patients could be that they



(**a**) A healthy brain at age 71

(**b**) An AD brain at age 73

Figure 2.9: Examples of a CN brain and an AD brain. Regions largely affected by AD have been marked out with red arrows. Images are taken from ADNI [5].

forget a conversation occurring a day ago. In addition, the lateral temporal cortex, particularly the medial and superior temporal gyrus, is also largely affected by AD [89], while the frontal cortex supporting executive functions witnesses less AD effect. Examples of a healthy brain and an AD brain are shown in Figure 2.9. However, the mechanisms of AD are yet not fully understood, and more data are required to uncover its secret.

Parkinson's disease (PD) is the second most common neurodegenerative disease. Individuals with PD can suffer from difficulty with walking, talking, balance and coordination. Although the cause of PD is still to be fully revealed, studies have found that PD is associated with abnormal (accelerated) volume decrease in several brain areas such as hippocampus, thalamus and anterior cingulate.

In general, people with neurodegenerative diseases exhibit more severe cognitive declines accompanied by the accelerated atrophy of particular brain areas. The difficulty with diagnosing these diseases lies in the fact that they mostly occur in later life and thus are tangled with the effects of normal ageing or even with each other.

2.5 Datasets

In this thesis we use four public medical datasets to validate the proposed approaches.

2.5.1 Ischemic Stroke Lesion Segmentation (ISLES) 2015 Challenge

Ischemic Stroke Lesion Segmentation (ISLES) [20] dataset consists of 28 subjects that are imaged in T1w, T2w, FLAIR, and DWI sequences. The MRI sequences were skull-stripped using BET2 [90], resampled to an isotropic spacing of $1\ mm^3$, and rigidly co-registered to the FLAIR sequences using the Elastix toolbox [91]. The segmentation annotations were labelled by experts on FLAIR sequences. All volumes were taken from subjects suffering from sub-acute ischemic stroke lesions. The ISLES dataset is used in Chapter 4 for evaluating our pseudo healthy synthesis model.

2.5.2 Multimodal Brain tumour Segmentation (BraTS) 2018 Challenge

Multimodal Brain tumour Segmentation (BraTS) [4] dataset consists of patients with high and low-grade gliomas. The MRI images were obtained from different centres using MR scanners from different vendors with different scanning settings. The subjects were imaged in T1w, T1c, T2w and FLAIR sequences. All volumes were skull-stripped [92], re-sampled to $1\ mm^3$ resolution, and rigidly co-registered to the T1c sequences using the ITK tool [93]. The BraTS dataset is used in Chapter 4 for evaluating our pseudo healthy synthesis model.

2.5.3 Cambridge Centre for Ageing and Neuroscience (Cam-CAN)

Cambridge Centre for Ageing and Neuroscience [2] (Cam-CAN) is a cross-sectional dataset containing normal subjects aged from 18 to 87. This dataset is large-scale with approximately 700 subjects and contains multi-modal data, including MRI. All MRI images were collected at the Medical Research Council (UK) Cognition and Brain Sciences Unit (MRC-CBSU) using a 3 T Siemens TIM Trio scanner. In this thesis, Data pre-processing were performed on the Cam-CAN data before using them for experiments. The data were first skull-stripped

using DeepBrain[1] and then linearly registered to Montreal Neurosciences Institute (MNI) 152 space using FSL-FLIRT [94]. The Cam-CAN dataset is used in Chapter 4 for learning a pathology-free distribution and in Chapter 5 for learning to simulate brain ageing progression.

2.5.4 Alzheimer's Disease Neuroimaging Initiative (ADNI)

Alzheimer's Disease Neuroimaging Initiative [5] (ADNI) is a large-scale dataset containing longitudinal studies. The subjects are cognitively normal (CN), mildly cognitive impaired (MCI) or with Alzheimer's Disease (AD). This dataset provides multiple types of information, including clinical data, MR images (processed and unprocessed T1-w and T2-w images and functional MRI), PET images, etc. In this thesis, brain MRI data are used for experiments after preprocessing. Similar to Cam-CAN data, the ADNI data are first skull-stripped and then linearly registered to MNI 152 space. The ADNI dataset is used in Chapter 5 for learning to simulate brain ageing progression considering health state and evaluating our model using longitudinal data.

2.6 Data preprocessing

In this thesis, several techniques have been used to preprocess the brain MRI data. Note BraTS and ISLES were preprocessed by their providers, and we only preprocess Cam-CAN and ADNI data in this tehsis.

2.6.1 Brain extraction (skull stripping)

Brain extraction refers to the process of accurately segmenting brain from non-brain tissue, e.g. skull. This is beneficial to many applications, where we only want to focus on the brain. In this thesis, we focus on the change of a brain from pathological to 'healthy' (Chapter 4) and the change of a brain with age (Chapter 5). As such, we choose to first remove non-brain tissues from MRI images.

[1] https://github.com/iitzco/deepbrain

The brain extraction (or skull stripping) technique used in this thesis is Brain Extraction Tool (BET) [95, 90]. BET first estimates a rough brain/non-brain threshold based on intensity histogram. The threshold is then used to measure the centre of gravity and mean radius of brain. After that, the initial surface model is created as a tessellated sphere, which is then updated iteratively calculating within-surface vertex spacing, surface smoothness control, and brain surface selection term. The final brain is obtained using the brain surface model. For more details please refer to [95, 90]. In this thesis, we use FSL-BET, i.e. BET algorithm provided by the FSL software.

We also use DeepBrain, which is a Python package for medical image processing, available at `https://github.com/iitzco/deepbrain`. The brain extraction function of DeepBrain is achieved by training a U-Net model on a variety of manual-verified skull-stripping datasets. DeepBrain is slightly faster than FSL-BET as it enables the use of GPU for computation.

2.6.2 Registration

Image registration refers to aligning multiple images, in order to compare and observe the spatial features of anatomy across images [96]. In this thesis, we choose to register the skull-stripped brain MRIs to a common space, i.e. the MNI 152 space, in order to better observe changes in synthetic brain images.

Registration can be classified into two types: linear/rigid registration and non-linear/non-rigid registration. Linear registration involves six-parametric rigid transformation, i.e. rotation and translation on x, y and z coordinate axes, or 12-parametric affine transformation, i.e. rotation, translation, scaling, and shearing on x, y, and z coordinate axes. Non-linear registration involves non-linear transformation such as local deformations. In this thesis, we use linear registration to pre-process brain MRIs, as it maintains the anatomical structure of brains. Specifically, we use FSL-FLIRT [94] with 12-parametric affine transformation, with the following parameter settings: *bins='256', cost=' corratio' , searchrx=' -90 90', searchry=' -90 90', searchrz=' -90 90', dof=' 12', interp=' trilinear'*.

2.7 Summary

This chapter provided background material Magnetic Resonance Imaging (MRI), as well as different MR sequences that are widely used. A brief introduction of brain anatomy and the brain ageing process was also presented. Then a description of two types of brain pathology, ischemic stroke and brain tumours, was provided. Finally, datasets used throughout this thesis were introduced. The next chapter focuses on the technical part, as well as a literature review of recent image synthesis with deep learning methods.

Chapter 3
Technical background

3.1 Machine learning

Artificial intelligence (AI) has been a thriving research field in recent years with a large number of applications [97, 17, 98]. The aim of AI is to create artificial beings (*e.g.* computers) that can simulate human intelligence and solve tasks that are either tiresome or difficult for human beings [98]. It is ironic that some abstract and formal tasks that are difficult for human beings turn out to be among the easiest for computers, such as playing chess and complex calculations. By contrast, some tasks that are naturally easy for human beings are difficult for computers, such as conversation and recognizing objects [97]. Dealing with everyday life requires a massive amount of knowledge about the world, and most of this knowledge is subjective and intuitive. Incorporation of this informal information into computers or any other intelligent machines becomes the key to realizing artificial intelligence [97].

The earliest attempts to achieve AI sought to hard-code informal knowledge about the world into formal computer languages. This is called the *knowledge base* approach. A famous example is Cyc [99], which is an inference engine with a large database of statements called CycL. These linguistic statements are manually entered by human experts, which is tedious and time-consuming labour. Describing the world with enough complexity and accuracy using formal language that can be understood by computers has always been a troublesome task. There is always something missing or some situations not considered, resulting in failures of AI system even in a seemingly simple task. For example, the Cyc system failed to understand a story about a man called Fred shaving in the morning [97]. It detected an inconsistency in this story: people do not have electrical parts, and Fred was holding an electric razor while shaving. Therefore, it threw a question that sounds strange: was Fred still a person while he was shaving? Since it is nearly impossible to hard-code all knowledge

into AI systems, one may ask if it is probable to let these AI systems acquire knowledge by themselves. This approach is known as *machine learning*.

Machine learning is a subfield of AI, where algorithms automatically *learn* to make predictions or decisions without being explicitly programmed to do so. A typical machine learning scenario involves a dataset of N pairs of data points $\{(x, y)\}_1^N$, where x is a sample from the input distribution: $x \in X$, y is the corresponding target output from the output distribution: $y \in Y$, and the aim is to find a mapping function: $f : X \to Y$. A simple example of such mapping functions could be using logistic regression to determine whether to recommend cesarean delivery or not [100]. Early machine learning algorithms such as logistic regression and naïve Bayes do not make predictions from *raw data* such as MRI scans but from hand-crafted *features* of the input data, where a feature could be a piece of information such as the presence of a uterine scar. As a result, the performance of these simple machine learning approaches largely depends on the quality of the features or the representation of the data.

However, for some tasks, it could be difficult to decide which features should be used. For example, if we need to make a program to detect cats in photos, what features should we use? Clearly, a cat should have two eyes, two ears, one mouth, four claws, and a tail. But if we simply use these features to detect cats, may get a number of dogs as they also possess the same features. Furthermore, how do we detect say eyes of cats? It is hard to say what cat eyes are like in pixel values. we know that cats' eyes are (nearly) round-shaped in geometry, but what if a cat in a photo happened to close her eyes? One solution to overcome the difficulty of designing hand-crafted features is to use machine learning to discover the features. One approach to extract features from raw data is to use *Artificial Neural Networks* (ANN).

Inspired by biological brains, ANN consist of artificial *neurons* and *edges*. Each neuron takes inputs and produces an output that can be provided to other neurons, whilst each edge connects two neurons by transmitting the output of one neuron to another and is assigned with a weight representing its relative importance. The neurons and edges are normally arranged into a series of stacked layers. When the number of the stacked layers is beyond 2, the resulting neural network is termed *deep neural network* (DNN), and the corresponding learning process is termed *deep learning*.

Recent years have witnessed great success of deep learning in many tasks, including image recognition [101, 102], speech recognition [103], drug discovery [104], particle accelerator data analysis [105], and medical image analysis [18]. The success of deep learning in the recent decade is largely due to the rapid development of computational power and much more availability of digital data. Deep learning can be classified into *supervised learning*, *semi-supervised learning* and *unsupervised learning*, based on the availability of input-output pairs in the training dataset and the type of pairing. Taking advantage of the abundance of data, the most common form of deep learning is *supervised learning*, where each input sample has a corresponding target output (also called a target label). By contrast, when there are no pairs of input and output samples, the learning process will learn patterns from the input data distribution and is named *unsupervised learning*. *Semi-supervised learning* concerns a mixed situation where only a part of input data have target labels.

Although supervised learning has achieved considerable success in many fields, it requires a large amount of training data with annotated labels. In some fields, *e.g.* medical imaging analysis, the image data and corresponding labels can be very expensive and time-consuming to acquire. As a result, medical datasets are relatively small, with the size of thousands or even less than a hundred, while natural image datasets normally contain millions of images. Supervised learning approaches that work well on larger natural datasets may perform worse on the limited medical datasets. One approach to alleviate such data scarcity is image synthesis. In the following sections, we will briefly introduce and discuss image synthesis models and then review recent works in pseudo healthy synthesis and brain ageing synthesis.

3.2 Generative Adversarial Networks

In the machine learning context, a *generative model* aims to capture the distribution of a training dataset in order to generate new data with variations. Normally, it is unrealistic to learn the exact distribution of the training data. Therefore, most existing approaches try to model a distribution that is as close to the true data distribution as possible. One great example could be Gaussian Mixture Models (GMMs), where the underlying data distribution is approximated using a weighted sum of K Gaussian distributions. For one-dimensional

data x, the probability functions can be represented as:

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

where ϕ_i, μ_i and σ_i are the weight, mean and variance of the i-th Gaussian distribution, and $\sum_{n=1}^{K} \phi_i = 1$. When K is known, we can use the Expectation Maximisation (EM) algorithm [106] to estimate the parameters of the model. During inference time, new data are synthesised by simply sampling from the learned mixed Gaussian distributions. Although classical approaches such as GMMs work well for low-dimensional data, they fail to represent the distributions of very high dimensional data such as MRI images. Techniques such as dimensionality reduction could be used to help with dealing with high dimensional data.

The rise of deep learning has offered an alternative solution, where deep neural networks are used to approximate the underlying complex distributions of training data. In this thesis, we develop the proposed models in Chapter 4 and 5 based on a popular deep generative model, *Generative Adversarial Networks* (GANs) [107].

3.2.1 Formulation of GANs

Generative adversarial networks (GANs) [107] are deep generative models that are trained within an adversarial process. A typical GAN consists of two components: a generative model *G* that aims to characterise the data distribution, and a discriminator model *D* that measures the likelihood of a sample coming from the real data distribution or produced by the generative model. These two sub-networks are trained simultaneously and adversarially, which is analogous to a *min-max* two-player game. In the optimal case, an equilibrium can be achieved with *G* capturing the training data distribution and *D* predicting $\frac{1}{2}$ for all samples. A schematic of GAN is shown in Figure 3.1.

Let us define the data distribution as $p_{data}(x)$ and the prior distribution of the random variable as $p_z(z)$, where the prior distribution $p_z(z)$ is typically a multivariate Gaussian $p_z(z) \sim \mathcal{N}(0, I)$. For vanilla GAN [107], the generator transforms a random variable z into a data space $G(z)$, and the aim is to encourage the generator's distribution p_g match the data distri-

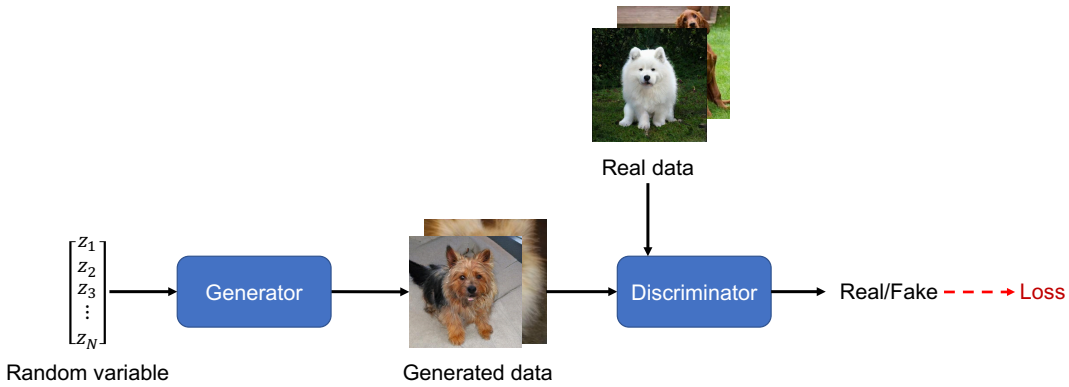


Figure 3.1: Schematic of a Generative Adversarial Network (GAN). The generator takes as input a random variable sampled from a known distribution and tries to produce output data; the discriminator classifies between real and generated data. The generator and discriminator are trained adversarially, with the discriminator trained to tell apart real and fake data and the generator trained to produce data that can be misclassified as real by the discriminator.

bution p_{data}. The discriminator outputs a single scalar $D(x)$ which represents the probability of a sample x coming from the true data distribution p_{data} rather than from p_g. The generator and discriminator are deep neural networks that can be trained with a two-player *min-max* game. The loss function is [108]:

$$\min_G \max_D V(G, D) = \min_G \max_D \mathbb{E}_{x \sim p_{data}(x)} \left[\log D(x)\right] + \mathbb{E}_{z \sim p_z(z)} \left[\log(1 - D(G(z))\right]. \quad (3.2)$$

The training is performed in a successive procedure. When training the generator, the weights of the discriminator are frozen and not updated. Similarly, the weights of the generator are frozen when updating the discriminator. If G is fixed, the optimal discriminator is:

$$D^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}. \quad (3.3)$$

If an optimal discriminator in Eq. 3.2 is considered, this *min-max* game can be reformulated

as:

$$
\begin{aligned}
V(G, D^*) &= \mathbb{E}_{x\sim p_{data}(x)}\left[\log D^*(x)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log(1 - D^*(G(z))\right] \\
&= \mathbb{E}_{x\sim p_{data}(x)}\left[\log D^*(x)\right] + \mathbb{E}_{x\sim p_g(x)}\left[\log(1 - D^*(x)\right] \\
&= \mathbb{E}_{x\sim p_{data}(x)}\left[\log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}\right] + \mathbb{E}_{x\sim p_g(x)}\left[\log \frac{p_g(x)}{p_{data}(x) + p_g(x)}\right] \\
&= \mathbb{E}_{x\sim p_{data}(x)}\left[-\log(2) + \log \frac{p_{data}(x)}{(p_{data}(x) + p_g(x))/2}\right] \\
&\quad + \mathbb{E}_{x\sim p_g(x)}\left[-\log(2) + \log \frac{p_g(x)}{(p_{data}(x) + p_g(x))/2}\right] \\
&= -\log(4) + D_{KL}\left(p_{data}\middle|\middle|\frac{p_{data}}{p_{data} + p_g}\right) + D_{KL}\left(p_{data}\middle|\middle|\frac{p_g}{p_{data} + p_g}\right),
\end{aligned}
\tag{3.4}
$$

where D_{KL} is the Kullback-Leibler (KL) divergence. The two KL terms in the previous equation can also be expressed by the Jensen-Shannon divergence (JSD) between p_g and p_{data}:

$$
V(G, D^*) = -\log(4) + 2 \cdot JSD(p_{data}||p_g). \tag{3.5}
$$

This means that when the discriminator is optimal, the generator is trained to minimise the JSD between the generated distribution and the true data distribution. By definition, the JSD is symmetric, *i.e.* $JSD(p_{data}||p_g) = JSD(p_g||p_{data})$, and non-negative. Specifically, $JSD(p_{data}||p_g)$ is zero if, and only if, $p_{data} = p_g$. Hence, the training convergence of GAN is achieved when the generator can produce a probability distribution equal to the true data distribution, $p_g = p_{data}$. In this ideal case, the discriminator is unable to tell apart true and generated data, and $D(x) = \frac{1}{2}$.

3.2.2 Issues with GANs

However, in practice the training of GANs is challenging, and the global optimality (*i.e.* $p_g = p_{data}$) is difficult to achieve [108]. Common problems with GANs include *mode collapse*, *non-convergence* and *vanishing gradients*.

When mode collapse happens, the generated distribution only contains a single or several modes of the true distribution. In mode collapse, the generator produces a limited variety of

samples that are realistic enough to cheat the discriminator. Since the discriminator is cheated by these generated samples, it will not back-propagate gradients to correct the generator. In short, the generator is stuck in local optimality.

The generator improves with training while the discriminator performance becomes worse as it is harder to classify between real and fake samples. Imagine the generator is trained perfectly, then discriminator outputs $\frac{1}{2}$ for everything, which means the discriminator is flipping a coin for every sample. This raises an issue for the convergence of GAN: if the generator continues training after the discriminator offers completely random predictions, the performance of the generator could drop because it trains on junk feedback. As a result, the GAN training oscillates, and convergence is often a fleeting, rather than stable, state.

Recall that when the discriminator is optimal, the objective function becomes Eq. 3.5, which aims to minimise the JSD between distributions of true data p_{data} and generated data p_g. However, if the overlap between p_g and p_{data} is little or even does not exist, the JSD will saturate, resulting in a negligible gradient being back-propagated to the generator [108]. In this case, the generator gradients vanish, and the generator learns nothing. When the discriminator trains too successfully and becomes optimal before p_g learns to match p_{data} to some extent, the vanishing gradient problem is highly likely to occur.

In order to solve the problems discussed above, considerable effort has been placed on proposed different approaches to improve the training of GANs.

3.2.3 Different ways to improve GAN training

As mentioned in the previous section, training GANs is difficult and unstable. A great amount of research has focused on improving the training stability of GANs. In this thesis, we take advantage of these works to improve and stabilise the training of our models. Below we give a brief introduction of recent works that focus on improving GAN training.

As an example, Deep Convolutional GAN (DCGAN) has been proposed in [109] which improves the vanilla GAN by modifying the network architecture. Specifically, the authors replaced the fully-connected layers of G and D with deep convolutional networks and also

replaced pooling layers with strided convolutions, which improved the training stability and image size and quality.

Furthermore, many developments on improving GAN training attempted to replace the original binary cross-entropy training loss. For example, Least Squared GAN (LSGAN) has been proposed in [110] which uses Least-squared (LS) loss as the training loss in [110] to provide smooth and non-saturating gradients. With the LS training loss, the generator trains to minimise the squared distance between generated samples and decision boundary, and hence the samples that are far away from the decision boundary are heavily penalised. The objective functions of the LSGAN are as below:

$$\begin{aligned} \min_D V_{LSGAN}(D) &= \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(D(x)-b)^2] + \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z))-a)^2], \\ \min_G V_{LSGAN}(G) &= \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z))-c)^2], \end{aligned} \tag{3.6}$$

where a and b are the labels for fake and real data, and c is the value that G wants D to predict for fake data. Common choices for these parameters include $a = -1, b = 1, c = 0$ or $a = -1, b = 1, c = 1$.

In [108], the authors provided a convincing theoretical analysis of the training of GANs, pointing out problems with using JSD as GAN loss. To solve these problems, the same authors proposed to replace the JSD with the Wasserstein distance as the training loss in [111]. The Wasserstein distance is also called *Earth-Mover* (EM) distance, which measures the least cost to transport one distribution to another distribution by moving the ”mass”, where cost is mass times transport distance. The advantage of the Wasserstein distance is that it is differentiable almost everywhere. Hence, the Wasserstein distance can provide stable gradients in cases where the JSD is locally saturated. This has been shown theoretically and empirically to help solve the vanishing gradient problem. Nevertheless, in order to compute the Wasserstein distance in a tractable way, the discriminator functions $f(x)$ are required to be 1-Lipschitz. [1] To enforce the Lipschitz constraint, the authors constrained the weights by simply clipping all weights to a fixed range, *e.g.* $[-0.01, 0.01]$, after each update step.

[1] No lines connecting any two points on the function have a gradient greater than 1: $|f(x_1) - f(x_2)| \leq |x_1 - x_2|$.

However, the authors admitted that clipping weight is not a good way to enforce the Lipschitz property. When the clipping parameter is small, it can lead to the vanishing gradient problem. By contrast, a large clipping parameter could result in gradients exploding. When the clipping range is selected well, the resulting Wasserstein Generative Adversarial Network (WGAN) proves to be stable and effective. The objective functions of WGAN are:

$$\begin{aligned} &\min_D V_{WGAN}(D) = \mathbb{E}_{z\sim p_z}[D(G(z))] - \mathbb{E}_{x\sim p_{data}}[D(x)], \\ &\min_G V_{WGAP}(G) = -\mathbb{E}_{z\sim p_z}[D(G(z))]. \end{aligned} \tag{3.7}$$

To solve the issues of weight clipping, *gradient penalty* has been proposed in [112] as a 'soft' way to enforce Lipschitz constraint, by applying a penalty on the gradient norm of the discriminator[2] output with respect to its input. To penalise the gradient norm, they omitted Batch Normalization layers in the discriminator. The penalty of gradient norm naturally pushes the discriminator towards 1-Lipschitz and hence enables stable training. The resulting Wasserstein GAN with Gradient Penalty (WGAN-GP) has been shown to perform better than the original WGAN and train stably for a variety of GAN architectures without careful hyperparameter tuning. The objective functions of the WGAN-GP are:

$$\begin{aligned} &\min_D V_{WGAN-GP}(D) = \mathbb{E}_{z\sim p_z}[D(G(z))] - \mathbb{E}_{x\sim p_{data}}[D(x)] + \lambda \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}[(|| \bigtriangledown_{\hat{x}} D(\hat{x})||_2 - 1)^2], \\ &\min_G V_{WGAN-GP}(G) = -\mathbb{E}_{z\sim p_z}[D(G(z))], \end{aligned} \tag{3.8}$$

where $\hat{x} = \epsilon x + (1-\epsilon)G(z)$, $\epsilon \sim U[0,1]$, and λ is the weight for the gradient penalty term, set as 10 in [112].

Some works also use auto-encoders as discriminators for GAN training, such as *Energy-Based Generative Adversarial Networks* (EBGAN) [113] and *Boundary Equilibrium Generative Adversarial Networks* (BEGAN) [114]. These auto-encoder like discriminators first extract latent features of input data and reconstruct the input using these features. The idea is that poorly generated images can result in large reconstruction errors since they miss the

[2] In [112] the discriminator is called as *critic*.

latent features that are required by the decoder for reconstruction. In contrast, a realistically-looking image should have low reconstruction errors. Hence, these approaches measured the divergence between reconstruction losses of real and fake samples, and the generator aimed to minimise the divergence either measured in energy-based metric [113] or in Wasserstein distance [114]. The objective functions of EBGAN are:

$$\begin{aligned} &\min_D V_{EBGAN}(D) = \mathbb{E}_{x\sim p_{data}}[D(x)] + \mathbb{E}_{z\sim p_z}[m - D(G(z)]^+, \\ &\min_G V_{EBGAN}(G) = \mathbb{E}_{z\sim p_z}[D(G(z))], \end{aligned} \tag{3.9}$$

where $[\cdot]^+ = max(0, \cdot)$, m is a pre-defined positive margin, and $D(x)$ is the reconstruction loss, *e.g. Mean Squared Error* (MSE) in [113]: $D(x) = MSE(Dec(Enc(x)), x)$. From these objective functions, we can see that the auto-encoding Discriminator to achieve low reconstruction loss for real samples. Here the Discriminator is encouraged to reduce the reconstruction error for real data and increase the reconstruction error for fake data if is below margin m. By contrast, the Generator is encouraged to produce samples that achieve low reconstruction errors.

The objective functions of BEGAN are:

$$\begin{aligned} &\min_D V_{BEGAN}(D) = \mathbb{E}_{x\sim p_{data}}[D(x)] - k_t \cdot \mathbb{E}_{z\sim p_z}[D(G(z)], \\ &\min_G V_{BEGAN}(G) = \mathbb{E}_{z\sim p_z}[D(G(z))], \\ &k_{t+1} = k_t + \lambda_k(\gamma\mathbb{E}_{x\sim p_{data}}[D(x)] - \mathbb{E}_{z\sim p_z}[D(G(z)]), \end{aligned} \tag{3.10}$$

where k_t is involved to maintain equilibrium between generator and discriminator, λ_k is the learning rate for k, γ is called *diversity ratio* which controls the trade-off between image diversity and quality, and $D(x)$ is the auto-encoding reconstruction loss. Similar to EBGAN (Eq. 3.10), the discriminator of BEGAN has two goals: achieving low reconstruction loss for real samples and high reconstruction loss for fake data, and the generator aims to lower the fake reconstruction loss by improving the quality of fake data. Regardless of the use of k_t term, we could observe similarities between Eq. 3.9 and the objective functions of WGAN (Eq. 3.7). The difference is that BEGAN measures the Wasserstein distance between the reconstruction loss distributions of real and fake samples, while WGAN leverages the

Wasserstein distance of data samples. Besides, BEGAN does not require Lipschitz constraints on the discriminator.

Since images are complex and high dimensional, instead of directly modelling whole images, prior studies [115, 116] proposed to first model low-resolution images and then gradually increase the image resolution to learn more complex data distribution. To be specific, Laplacian Generative Adversarial Networks (LAPGAN) [115] combines a Laplacian pyramid representation [117] with GANs by using a set of generative models to capture the distributions of natural images at different levels of the Laplacian pyramid. Similarly, Progressive Growing GANs (PGGAN) [116] started from training on low-resolution images and then added layers to existing neural networks as image resolution increased in a smooth way, *i.e.* fade in new layers gradually. In short, these multi-scale approaches first learn a global abstract of natural images and then fine-tune the details.

In summary, most techniques to improve the training of GANs focused on alleviating unstable training issues, by modifying architectures, by replacing the original objective functions, or by optimizing training strategies. Other techniques that help GAN training include spectral normalization [118] which constrains the discriminator to be Lipschitz continuous, memory replay [119] which presents previously generated images to the discriminator to prevent forgetting, and data augmentation [120, 121] which helps prevent over-fitting of both generator and discriminator.

3.2.4 GAN variants

The original GAN [107] generates data from random latent vectors but does not control what data to generate. To enable more fine-grained control of what contents to generate, different variants of GANs have been proposed. In this thesis, we also modified the structure of GANs to suit our tasks. Below we provide a brief review of GAN variants.

Conditional GAN: Conditional Generative Adversarial Networks (conditional GANs, or cGANs) are a popular variant of GANs. First proposed in [122], conditional GANs condition the Generator and Discriminator on some extra information y. The objective function

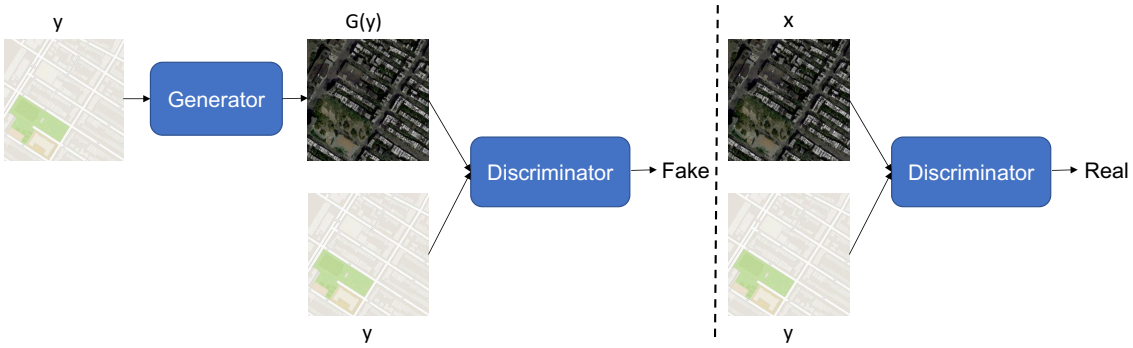


Figure 3.2: A schematic of pix2pix GAN [6]. Here the aim is to learn a mapping from a map to an aerial photo. The discriminator learns to classify between a fake pair consisting of a generated aerial photo $G(y)$ and the input map y and a real pair consisting of a ground-truth aerial photo x and the input map y.

of conditional GANs in [122] is:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log(x|y)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z|y)))], \tag{3.11}$$

where x is the real data sample, z is the random latent code, and y is the extra information on which the Generator and Discriminator are conditioned. Depending on different types of information y, conditional GANs can be used to generate data samples conditioned on class labels [123, 124], text [125, 126, 127], and bounding box [128].

Conditional GANs can also be used in the context of image-to-image translation by taking the extra information y in the form of images. Specifically, pix2pix GAN has been proposed in [6] (see Figure 3.2) that aims to learn a mapping from an observed image y to output image $G(y)$, for instance, from a map to an aerial photo.

For cGANs in [122], the Generator takes a random vector z as input and produces a data sample y (*e.g.* an image). By contrast, pix2pix GAN [6] do not involve a random noise z. Instead, the Generator of pix2pix GANs learns a mapping from a given image y to an output

image of desired properties. The objective function of pix2pix GAN is:

$$\begin{aligned}\min_G \max_D V(D,G) = &\mathbb{E}_{x\sim p_{data}, y\sim p_y}[\log D(x,y)] + \mathbb{E}_{y\sim p_y}[\log(1 - D(y, G(y)))] \\ &+ \lambda \mathbb{E}_{x\sim p_{data}, y\sim p_y}[||x - G(y)||_1],\end{aligned} \tag{3.12}$$

where an l_1 term is applied to encourage $G(y)$ to be close to x. Instead of mapping from a latent space to image space, pix2pix learns a mapping from an image space to another image space, which offers more controllability compared to original GANs and cGANs in [122]. Following the same spirit, pix2pixHD [129] adopts cGANs and uses feature matching loss for image synthesis and semantic manipulation.

Image-based cGANs like pix2pix require the pairing of image y and x, *e.g.* pairing of a map and an aerial photo. However, acquiring such pairs of image data often needs human effort and can be laborious and extremely difficult, which hinders the availability of paired training data. To solve this problem, CycleGAN has been proposed in [130] to learn unpaired image-to-image translation, as detailed below.

CycleGAN: As a special type of conditional GAN, CycleGAN was proposed in [130] with the aim of learning image-to-image mappings from unpaired data. Concurrently, DiscoGAN [131] and DualGAN [132] were proposed independently, sharing nearly the same framework and spirit as CycleGAN. Given a set of images in the domain X and a different set in the domain Y, CycleGAN learns a mapping G from X to Y, such that the output $\hat{y} = G(x)$, $x \in X$, is indistinguishable from images in the Y domain, $y \in Y$. This is learnt by adversarial training with a discriminator to classify between $\hat{y}$ and y. However, such translation alone does not guarantee a meaningful pairing between X and Y, and there are many mappings G that can satisfy $G(x) \in Y$. For instance, the model can simply learn a mapping to the same sample $y \in Y$, no matter which x is given, which is known as mode collapse. In order to solve this problem, CycleGAN leverages the idea *cycle consistency*, *i.e.* the translated output $\hat{y} = G(x)$ should be able to be translated back to x. Hence, CycleGAN involves another mapping F from Y to X, such that $x \approx F(\hat{y}) = F(G(x))$. Here G and F should be inverse of each other, and the mappings should be bijections. Similarly, we can leverage a sample from Y domain, $y \in Y$, as input, and get a generated sample in X domain, $\hat{x} = F(y)$. Then we can translate $\hat{x}$ back to domain X, $\hat{y} \approx G(\hat{x}) = G(F(x))$. A schematic

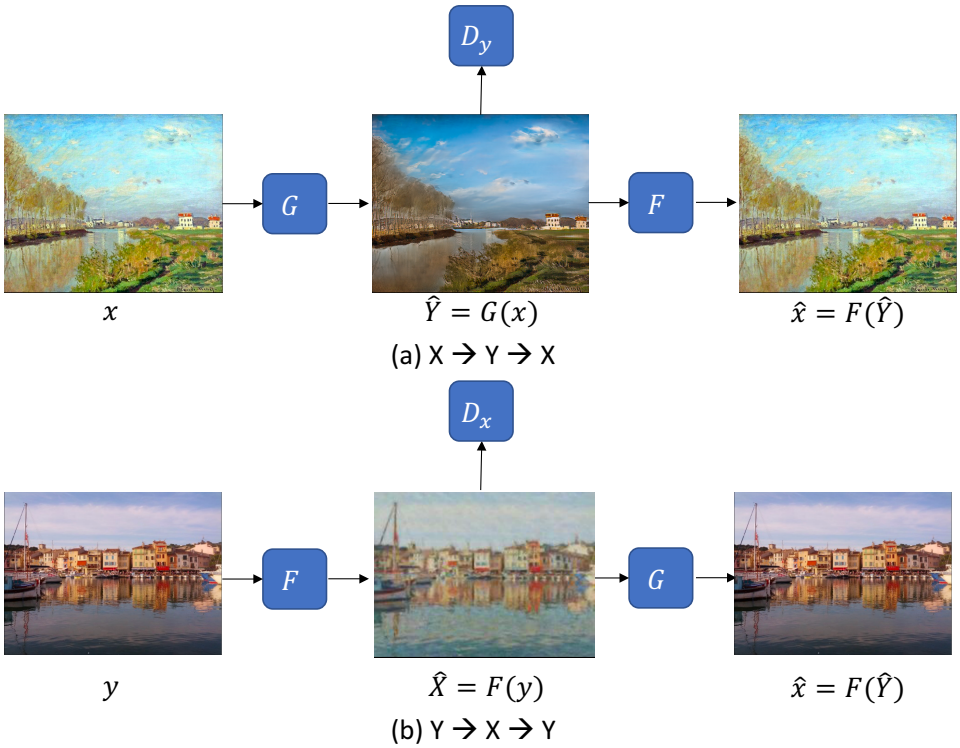


Figure 3.3: A schematic of CycleGAN. Here domain X represents Monet's style paintings, and domain Y denotes landscape photos. On top, a Monet painting is first translated to a photo and then translated back to the Monet domain; on bottom, a landscape painting is translated to Monet domain and then back to photo domain.

of CycleGAN is presented in Figure 3.3.

Let us denote the data distributions as $x \sim p_x(x)$ and $y \sim p_y(y)$. There are two Generators representing two mappings $G : X \to Y$ and $F : Y \to X$, and two discriminators D_X and D_Y, where D_X aims to classify between images $\{x\}$ and translated images $\{F(y)\}$, and D_Y classifies between images $\{y\}$ and $\{G(x)\}$. For the mapping function $G : X \to Y$, the adversarial objective function is:

$$\min_G \max_{D_Y} L_{GAN}(G, D_Y) = \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_Y(G(x)))], \tag{3.13}$$

where G tries to translate images x to $G(x)$ that are similar to images in domain Y, and D_Y aims to classify between translated images $G(x)$ and real images y of domain Y. Similarly,

the adversarial objective is:

$$\min_{F} \max_{D_X} L_{GAN}(F, D_x) = \mathbb{E}_{x \sim p_{data}(x)}[\log D_X(x)] + \mathbb{E}_{y \sim p_{data}(y)}[\log(1 - D_X(F(y)))], \tag{3.14}$$

where F tries to translate images y to $F(y)$ that are similar to images in domain X, and D_X aims to classify between translated images $F(y)$ and real images x of domain X.

In line with the adversarial losses, there is *cycle consistency loss*:

$$\min_{G,F} L_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[||F(G(x)) - x||_1] + \mathbb{E}_{y \sim p_{data(y)}}[||G(F(y)) - y||_1], \tag{3.15}$$

where the aim is to minimise the errors between input and reconstructions. Combining the above equations, the overall objective function of CycleGAN is:

$$L_{cycleGAN} = \min_{G} \max_{D_Y} L_{GAN}(G, D_Y) + \min_{F} \max_{D_X} L_{GAN}(F, D_x) + \lambda \min_{G,F} L_{cyc}(G, F), \tag{3.16}$$

where λ controls the relative importance of cycle consistency.

The advantage of CycleGAN is that it can capture specific characteristics of one image domain and translate these characteristics to another image domain without requiring any paired data. This lowers the data requirements to train an image-to-image translation model. As a result, CycleGAN has been widely used in image translation tasks. However, not all image domains contain the same amount of information. For example, if we want to translate a colourful image to a grey-scale image, there would be information loss since a colourful image naturally contains more information than its grey-scale counterpart. In this case, when we translate the greyish image back to colour, the colourful information has to either be invented by the model or hidden in the greyish image in a way that can be overlooked by the discriminator. This is also known as the ”one-to-many” problem.

GANs with an encoder: The original GAN only has the mapping from latent features to

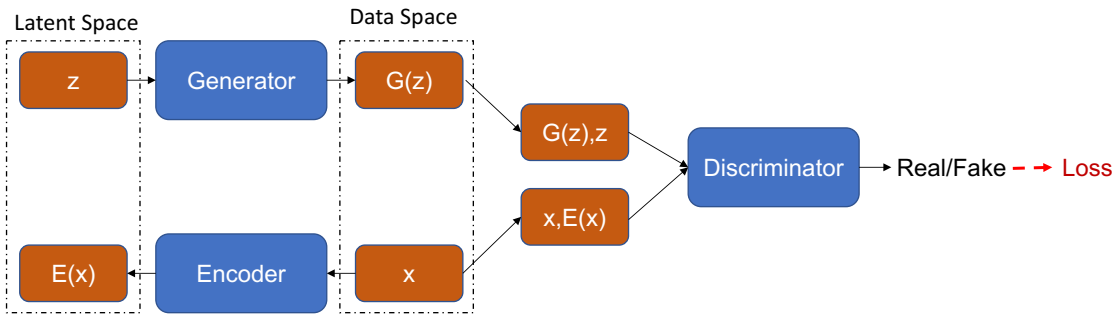


Figure 3.4: Schematic of BiGAN/ALI structure [7, 8]. The Generator is used to map a latent vector z to a generated data $G(z)$. The Encoder is used to map data x back to the latent space $E(x)$. The Discriminator takes as input a pair of data and its corresponding latent code. For real data, this pair is $\{x, E(x)\}$; for generated data, the pair is $\{G(z), z\}$.

data but does not learn the inverse mapping, *i.e.* projecting data back to the latent space. To solve this problem, Bidirectional GANs (BiGANs) was proposed in [7] by adding an encoder to GANs. The encoder was used to learn the inverse mapping, and the resulting learned feature representations were shown to be useful. Similarly, the Adversarially Learned Inference (ALI) model has been proposed in [8], which also applies an encoder to learn latent features. A schematic of ALI and BiGAN is shown in Figure 3.4. The objective functions of BiGAN/ALI are:

$$\begin{aligned}\min_{G,E}\max_{D} V_{BiGAN/ALI}(G,E,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x, E(x))] \\ + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z), z))],\end{aligned} \tag{3.17}$$

where x is a data sample, z is a latent code, E is the Encoder, D is the Discriminator, and G is the Generator. For BiGAN/ALI, the generator can be also viewed as a *decoder* since it maps from latent space to data space.

Similar to the idea of using an encoder to model the latent distribution, DeliGAN [133] uses a mixture of Gaussians to model the latent distribution and learns the mixture components by maximising the likelihood of generated samples under the data generating distribution. For encoding-decoding models, the decoder output, or the reconstruction, should be similar to the input, which regularises the latent space and encourages samples that are similar in data space to have similar latent vectors. Despite the ability to interpolate between real data samples,

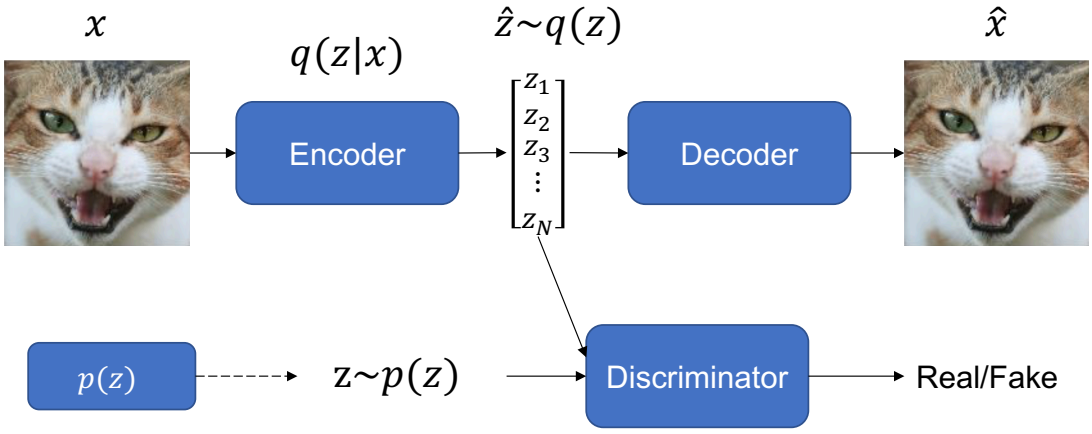


Figure 3.5: A schematic of AAE. The Encoder generates a latent vector $\hat{z}$ from a data sample x; the Discriminator classifies whether the generated vector $\hat{z}$ is from the prior distribution $p(z)$; and the Decoder obtains reconstruction $\hat{x}$ from the latent vector $\hat{z}$. The adversary encourages the generated vector $\hat{z}$ to be similar to random vectors from the prior distribution.

the image quality of the synthesised/reconstructed output of BiGAN/ALI is poor.

GANs with VAEs: There are a group of GAN variants that are based on both GANs and *Variational AutoEncoders* (VAEs) [134]. For instance, Adversarial Autoencoders (AAE) was proposed in [135], where a discriminator is applied on the latent space to tell if a latent vector is generated by the Encoder or randomly drawn from a prior distribution. A schematic of AAE is presented in Figure 3.5.

Similar to VAE, AAE first encodes a data sample x to a latent space, which is imposed on a prior distribution, and then decodes the latent sample $\hat{z}$ back to the data space. The difference is that VAE uses KL divergence to enforce the latent distribution $q(z)$ to be close to the prior $p(z)$, while AAE approaches this using adversarial training. The loss function for AAE is [135]:

$$\begin{aligned}\min_{E,Dec}\max_{D} V(E,E,Dis) &= \mathbb{E}_{x\sim p_{data}}|x - Dec(E(x))|_1 \\ &+ \lambda\{\mathbb{E}_{z\sim p_z}[\log D(z)] + \mathbb{E}_{x\sim p_{data}}[\log(1 - D(E(x)))]\},\end{aligned} \tag{3.18}$$

where E is the Encoder, Dec is the Decoder, the Discriminator is denoted as D to differ from Dec, and λ controls the relative importance of adversarial training.

While AAE leverages adversarial training to impose the latent space on the prior distribution, UNsupervised Image-to-image Translation network (UNIT) was proposed in [136] which combines GAN and VAE by using VAE to enforce a smooth latent space and using adversarial training to improve reconstruction quality. Furthermore, Adversarial Variational Bayes [137] utilises adversarial training for pairs of latent vector and data sample in the VAE framework, and the learned feature representations in the GAN discriminator is used as a bias for VAE reconstruction [138].

3.3 Medical image synthesis

In the previous section, several popular deep generative models were introduced. This section will give a brief introduction to medical image synthesis using traditional machine learning or deep learning models. Then two sub-topics will be discussed: *pseudo healthy synthesis* and *brain ageing synthesis* that are directly related to the proposed approaches.

Medical image synthesis represents approaches that aim to model mappings from some given source images or even latent vectors to the unknown target images [139]. For instance, the source images can be low-dose Computed Tomography (CT) images of relatively lower quality, and the target images can be the corresponding full-dose images of higher quality. Such image transformation could enhance image quality and reduce the scanning cost. Also, some approaches attempted to synthesise medical images of one modality, say CT, to another modality, say MRI T1. These approaches could capture helpful information in the source modality and present this information in the target modality without further scanning the subject.

Conventional approaches for medical image synthesis include dictionary learning [140] and random forest [141], which process hand-crafted medical image features manually selected by experts during the synthesis. However, the quality of their synthesised outputs heavily depends on these hand-crafted features, which often have limited capacities to represent the complex information in the medical images. In recent years, the rise of deep learning has offered a promising solution to this issue through automatically learning the powerful features for synthesising desired medical images.

Most deep learning-based medical image synthesis approaches adopted convolutional neural networks (CNN) based architectures. These approaches can be mainly classified into two classes based on their applications, *image quality enhancement* and *creation of unknown images*. As exemplified above, low-dose CT images were transformed to full-dose counterparts using CNN-based methods in [142, 143, 144]. Similarly, CNNs were utilised for super-resolution or image quality enhancement of MRI and PET images in [145, 146, 147, 148]. These approaches were mostly trained in a supervised manner where target outputs of higher quality or resolution are available.

Another group of medical synthesis methods tried to estimate unknown images. For instance, a 3D CNN was used to predict CT images from MRI images in [149]. Similarly, an *encoder-decoder* framework was proposed in [150] to translate images across multiple modalities . Other similar works include [151, 152, 153], all used CNN-based methods for image synthesis from one modality to another. However, these approaches required the pairing of source and target images, limiting their utility when data pairing is not easily available. To solve this issue, CycleGAN was used to learn mappings between medical modalities without using paired data in [154, 155, 156].

In some cases, it is nearly impossible to obtain a ground-truth target image. For example, suppose we have a medical image of a patient who has brain tumour, and the aim is to generate a medical image of this patient when he/she is healthy. This patient's 'healthy' image could allow us to check to which extent the brain tumour affects his/her brain. However, there is almost no way to obtain such a 'healthy' image in clinical practice. The synthesis of a 'healthy' version of pathological image is also known as *pseudo healthy synthesis*, which we will discuss in the following section.

3.3.1 Pseudo healthy synthesis

Pseudo healthy synthesis is a sub-field of medical image synthesis with the aim of creating a 'healthy' version of a pathological medical image. One challenge with pseudo healthy synthesis is the lack of ground-truth 'healthy' data, *i.e.* finding a pathological and a healthy image of the same subject remains difficult, since a subject cannot be healthy and diseased

at the same time. Another challenge is the preservation and evaluation of *subject identity*, *i.e.* how to ensure that the synthesised 'healthy' image belongs to the same subject as the original image when there is no ground-truth data. Below we briefly review the literature of pseudo healthy synthesis. More details are presented in Chapter 4.

Non-deep learning based methods: Prior to deep learning, conventional machine learning approaches to this task focused on learning manifolds between 'healthy' and 'diseased' local regions at the patch [9, 157] or voxel [158, 159] level. In [9], the task was defined as generating a target image S in the desired domain given a source image I in the source modality. To achieve this, the authors first constructed a dataset, or dictionary, which contains N exemplar pairs of source and target images: $\mathcal{T} = \{(I_n, S_n)\}_{n=1}^{N}$, where I_n is a medical image in the source domain, S_n is an image in the target domain, and I_n and S_n are spatially aligned. They made an assumption that similarities between I and $\{I_n\}$ would lead to similarities between S and $\{S_n\}$. Hence, S was synthesised using I and the dictionary $\mathcal{T}$ through patch-based nearest neighbour search, *i.e.* finding the closet patch in $\{I_n\}$ for each patch of I and propagating the corresponding patch in $\{S_n\}$ to construct S. Specifically, they synthesised T2 brain MRI images with tumours from T1 brain MRI images and subtracted the resulting 'pseudo healthy' T2 images from the ground-truth T2 images to detect the tumour regions. Here, the synthesis of 'healthy' images is based on the premise that tumours are less visible in T1 images, and hence the transformation from T1 to T2 could reduce the intensity abnormality of tumours. Figure 3.6 presents the visual results taken from [9], comparing with a method termed 'Warped Atlas' [10].

Similarly, a dataset was constructed in [157] which consists of chest radiographs diagnosed as 'normal' by experts. For a given input chest radiographs, the proposed algorithm searched for the most similar patches within the normal dataset and then constructed a pseudo normal image using these patches. The resulting 'healthy' image was then subtracted from the input chest radiograph to detect lung nodules. A main difference between [157] and [9] is that the dictionary of [9] contains pairs of source and target images, while the dictionary of [157] only consists of target (normal) images. Instead of dictionary-based learning, [158, 159] used voxel-wise kernel regression to learn a direct mapping between healthy T1-w and FLAIR intensities within each voxel. Then the learnt regression model was used to synthesise pseudo

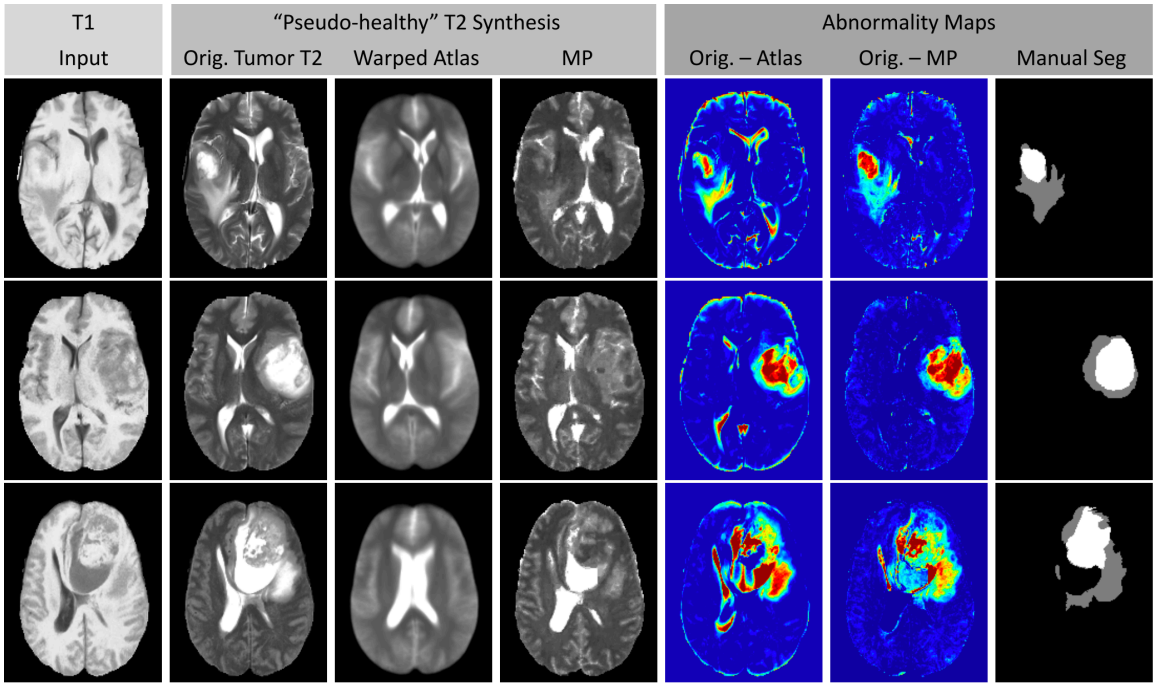


Figure 3.6: Visual results taken from [9]. Pseudo healthy T2 images were generated from T1 input and subtracted from original T2 images to obtain abnormality maps. 'Warped Atlas' is a comparison method [10]. MP (short for Modality Propagation) refers to the proposed method.

healthy FLAIR images from T1-w images. The success of pseudo healthy synthesis relied on the premise that the pathology is less visible in T1-w images, and the regression model is trained on healthy pairs of T1-w and FLAIR images.

However, these methods heavily depend on either the variation and size of the learnt dictionary [9, 157], or the capacity of the regression model [158, 159], which limits their ability to scale up to large medical images. Furthermore, pseudo healthy synthesis of [9, 158, 159] is also based on the premise that the target pathology is not dominant in some modalities but obviously visible in others, which may not be true for all kinds of pathology.

Autoencoder-based methods: This group of approaches scaled up patches to the image level using deep convolutional neural networks [160, 161, 162, 163, 164, 165, 166, 167, 11]. The main principle is to learn representations of the normal anatomy in an encoding-decoding manner, *i.e.* learning to compress and recover healthy data. There is an implicit

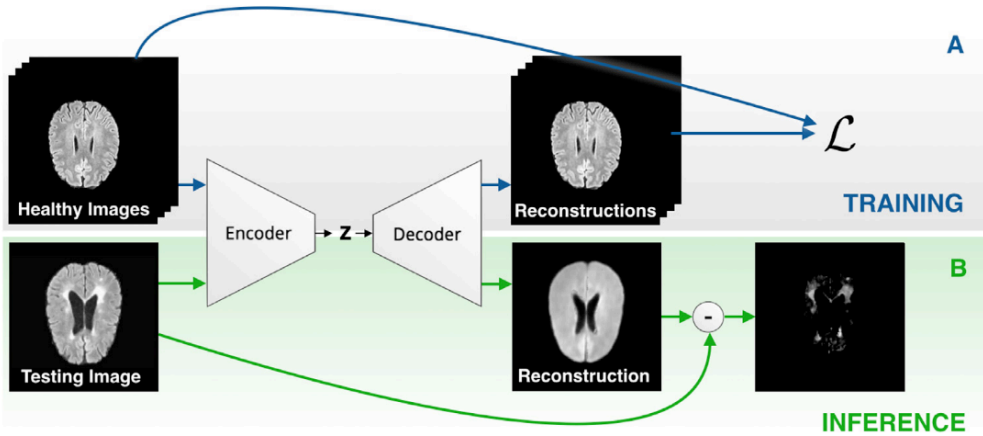


Figure 3.7: A schematic of autoencoder-based methods for pseudo healthy synthesis. A) Training a model on healthy data only; B) Pseudo healthy synthesis of a pathological image and pathology segmentation by subtracting the pseudo healthy image from original image. Figure taken from [11].

assumption that if the model is trained only on normal (healthy) data, it will only produce images within the normal distribution even when given pathological inputs. During inference, pseudo healthy images are generated by compressing and recovering pathological images, and these pseudo healthy images are then used for detection or segmentation of lesions. A schematic of these approaches is taken from [11] and presented in Figure 3.7.

Specifically, an Adversarial Autoencoder (AAE) based method was proposed in [160] to reconstruct healthy brain MRI images. Apart from the AAE loss function (see Eq. 3.18), they also proposed a regularization term to impose representation consistency:

$$L_{reg} = ||z_h - z'_h||^2, \tag{3.19}$$

where z_h is the latent representation of the healthy image x_h, and z'_h is the latent projection of the reconstructed image x'_h. During inference, the pseudo healthy images are obtained by simply feeding pathological images to the model, and the resulting images are then compared with the input to detect pathology.

Similarly, Gaussian Mixture Variational Auto-Encoder (GMVAE) was used in [161, 167] to learn the manifold of healthy images, and the pseudo healthy images were obtained using

the learnt manifold through a Maximum-A-Posterior (MAP) restoration model. Furthermore, VAE-based models were used in [164, 166] to capture the normal manifold, while Bayesian Autoencoder was adopted in [162]. Specially, a GAN was used in [163] to learn the manifold of normal data. However, as GAN only learnt the mapping from a latent vector z to data x: $G(z) = z \rightarrow x$ but did not learn the inverse mapping: $x \rightarrow z$, they proposed an iterative process to find the corresponding z for a given input image, by backpropagating a proposed loss measuring errors between $G(z)$ and x to z.

However, there is no loss to ensure that *subject identity*, *i.e.* the synthetic output and input images should belong to the same subject, will be maintained, and the generation of pseudo healthy synthesis is based on the assumption that these models only produce outputs within the normal distribution, but there is no explicit loss to guarantee this assumption to be true.

Generative models: The rise of Generative Adversarial Networks [107] (GANs) has provided new opportunities to pseudo healthy synthesis. The original GAN can learn a mapping from the latent space to the data space, but it does not learn the inverse mapping from data to latent space. Hence, it does not fit in the context of pseudo healthy synthesis where the goal is to transform a pathological image to its 'healthy' version. Although some iterative algorithms could be used to find the latent vector for a given image, it is not neat and computationally expensive. To solve this problem, variants of GANs have been proposed [122, 130, 6].

Specifically, Conditional GAN [122] in its simplest form can transform a pathological image to the healthy domain trained in an unpaired manner. However, there is no loss to enforce subject identity, and it is likely that the model learns to transform a pathological image of one subject to a healthy image of another subject. To help maintain identity, pix2pix GAN [6] used pairs of source and target images as input to the discriminator and adopted an l_1 reconstruction loss. Nevertheless, pix2pix GAN required access to paired data, which is difficult to acquire in the context of pseudo healthy synthesis.

Due to the lack of paired data, VA-GAN [12] adopted the form of a Conditional GAN, with a regularization loss between the input (pathological) and the output (pseudo healthy) to help preserve identity. Specifically, they treated the disease effect as a separate additive factor and formulated the image transformation problem as adding or subtracting a residual map,

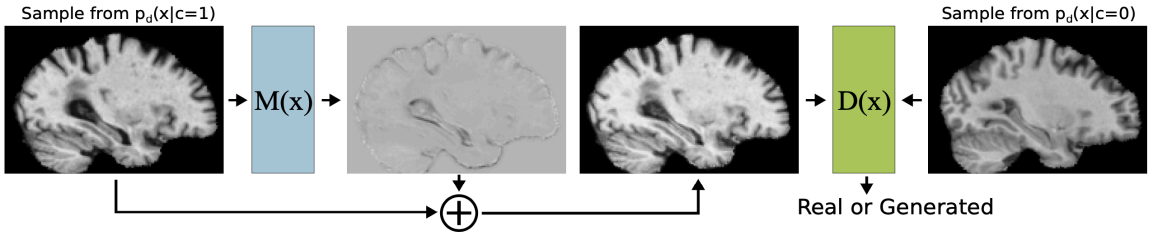


Figure 3.8: Schematic of VA-GAN. $M(x)$ refers to the generative network that produces the *disease effect map*, and $D(x)$ refers to the discriminator that judges if a given input is realistic and within the target domain. Image taken from [12].

termed *disease effect map*, to a healthy or pathological image. They defined two classes $c \in \{0, 1\}$, a source class and a target class. The distribution of images from class $c = 0$ was defined as $p_d(x|c = 0)$. Similarly, $p_d(x|c = 1)$ was the distribution of images from $c = 1$. Mathematically, the task is defined as:

$$\hat{y} = x + M(x), \tag{3.20}$$

where x refers to an image from the source class, say $c = 1$, $\hat{y}$ is the synthetic image which should be within the target class $c = 0$, and $M(x)$ is the additive map that contains the features distinguishing x from the other class. A schematic of VA-GAN is taken from [12] and presented in Figure 3.8.

VA-GAN adopted the Wasserstein GAN loss [111] for the adversarial training. The objective functions are defined as:

$$\begin{aligned} \min_D V_{VA-GAN}(D) &= \mathbb{E}_{x\sim p_d(x|c=1)}[D(x + M(x))] - \mathbb{E}_{y\sim p_d(y|c=0)}[D(y)] \\ &\quad + \lambda_1 \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}[(|| \bigtriangledown_{\hat{x}} D(\hat{x})||_2 - 1)^2], \\ \min_M V_{VA-GAN}(M) &= -\mathbb{E}_{x\sim p_d(x|c=1)}[D(x + M(x))] + \lambda_2 \mathbb{E}_{x\sim p_d(x|c=1)}||M(x)||_1, \end{aligned} \tag{3.21}$$

where $\hat{x} = \epsilon \cdot +(1 - \epsilon) \cdot y$, $\epsilon \sim U[0, 1]$, and λ_1 is the weight for the gradient penalty term; $||M(x)||_1$ is a regularisation on M that aims to encourage the identity consistency between x and $x + M(x)$, and λ_2 is the weight for the regularisation. λ_1 and λ_2 are both set as 10 in [12]. Notice that there are two main differences between Eq. 3.21 and the objective functions

of WGAN-GP (Eq. 3.8): the discriminator in Eq. 3.21 takes the sum of x and M_x as input, and there is a regularisation loss $L_{reg} = ||M_x||_1$ to help maintain subject identity.

However, there exists one potential problem for VA-GAN, *i.e.* the regularisation loss may conflict with the synthesis process. The focus of [12] was on the visual attribution of Alzheimer's Disease, where the disease effect is subtle and diffuse. But for other pathologies such as glioblastoma and ischemic stroke, the disease effect could be significant and localised, and thus it could be hard to balance the adversarial loss (which tries to make the input image 'healthy') and the regularisation loss (which tries to minimise the change).

Instead of an l_1 regularisation loss, CycleGAN [130] proposed the cycle consistency term to enforce identity preservation in the unpaired training. See Section 3.2.4 for more introduction of CycleGAN. CycleGAN has been used for pseudo-healthy synthesis [168, 169, 170, 171]. However, CycleGAN faces the *one-to-many* problem when one domain contains less information than the other [172]. Consider the case of pseudo healthy synthesis, if we transform a pathological image x_p to its healthy version $\tilde{x}_h$ perfectly, then the information of the pathology is lost. A question rises: *how can we transform $\tilde{x}_h$ back to $\hat{x}_p$ without knowing the pathology information?* To solve this problem, auxiliary information can be provided when synthesising $\hat{x}_p$, such as [169, 170] where they added a pathological residual to the healthy image to obtain the pathological image. But as discussed above, not all pathology can be considered as additive factors. Nevertheless, our approach in Chapter 4 aims to address the above shortcomings by disentangling images in pathological and anatomical factors. More details are presented in Chapter 4.

3.3.2 Brain ageing synthesis

As introduced in Chapter 2, our body goes through age-related changes when age increases [39]. Predicting future medical images based on current and past observations is another sub-field of medical image synthesis where the ground-truth paired data are extremely difficult to acquire. Consider the following question: *how will my face look like when we grow older, or more difficult, how will my brain look like when we get older?* One way to answer these questions could be to collect images of different ages and obtain a template representing

group-level changes. But an individual's ageing trajectory may differ from the group-average trajectory due to many factors. Thus, a method that can simulate subject-specific ageing trajectories could better solve the problem. Another question could be, what makes a good synthesis in the context of brain ageing synthesis. In general, a good synthetic brain image in the context of brain ageing should be *realistic*, *accurate in age*, and of the same *subject identity*. This section will briefly review literature that aims to model brain ageing. As face also undergoes age-related changes and there is more literature in face ageing synthesis, we also introduce recent works in the face ageing field.

Non-deep learning methods: Conventional approaches normally constructed group-average atlases to characterise the brain ageing progression. For instance, Zhang et al. [173] first used group-wise non-linear registration [174] and kernel regression to form an age-specific common space F_P which represented the average anatomy of infant brains of specific age points. Then they used the GLIRT group-wise non-linear registration [174] to align the templates in F_P to obtain a longitudinal common space F_L. Finally, a spatio-temporal atlas was constructed with patch-based dictionary learning and frequency domain sparse representation. Similarly, a group-wise non-rigid image registration method [175] and partial least squares regression is used in [176, 177] to construct a spatio-temporal reference model characterising brain ageing. A kernel regression method based on image dissimilarities was proposed in [178] to estimate brain images representative for different ages. In [179], the authors used pairwise non-rigid regression and kernel regression with adaptive kernel widths to construct a 4D spatio-temporal atlas of the developing brain. In [180], brain ageing was represented using linear mixed-effects modelling, and a brain image atlas was constructed using diffeomorphic registration parameterised by stationary velocity fields (SVFs). The constructed atlas represented normal ageing evolution and was used to separate the effect of normal ageing and the Alzheimer's disease. Following the same idea, a deformation based model conditioned on both age and disease was used to capture brain ageing considering the effect of Alzheimer's disease [181]. Similarly, approaches of [182, 183, 184] followed the similar manner of constructing average atlases using non-rigid registration, while biophysical models were used to simulate brain ageing in [185, 186]. However, these models relied on group-average atlases and thus could not model brain ageing specific to individuals.

Deep learning methods: Recent studies have tried to utilise deep generative models to simulate brain ageing trajectories. For example, a conditional GAN with Wasserstein training loss was used to generate a synthetically aged brain image given a baseline image [187]. Similarly, in [188, 189] the authors used conditional GANs to predict the evolution of white matter hyperintensities in brain MRI images. However, these approaches simply treated the problem as a domain transfer problem with two domains: *young* and *old*, and thus they could not explicitly synthesise brain images that are conditioned on some target ages. Although in [187] the generator was applied recursively to synthesise images of different ages, the *age accuracy* of the generated images was not encouraged by any loss. A deformation-based deep network was used to synthesise future brain images using longitudinal training data [190].

The authors of [191, 192] used a conditional adversarial autoencoder to simulate brain ageing trajectories, following recent work in face ageing generation [193], but they required longitudinal data to train the model and thus only covered several years of age spans. In [194], a GAN-based model was trained to add or remove atrophy patterns to brain images using image arithmetics. However, this approach was based on assumptions that atrophy patterns across ages could be modelled linearly and morphological changes were the same for all subjects. In [195], the authors utilised Variational Autoencoders to synthesise aged brain images, but they did not have control on the target ages, and the outputs appeared blurry. A structural casual model based on VAE was developed in [196] to generate brain aged images. However, this method did not provide quantitative evaluations of the synthetic results, and the resolution of the qualitative results was relatively low. In [197], the authors leveraged a VAE to disentangle spatial information from temporal progression and used the first layers of the trained VAE to improve brain age estimation. To summarise, most of previous works either built average atlases [173, 176, 198, 179, 180, 181] or required longitudinal training data [188, 189, 191, 192, 187]. Other did not evaluate morphological changes in detail and did not consider subject identity [194, 195, 196, 197].

To solve the aforementioned issues, in chapter 4, we propose a conditional adversarial approach that learns to synthesise subject-specific ageing brains without the need for longitudinal data.

3.3.3 Face ageing synthesis

Here we also briefly review some related work in face ageing generation, where the goal is to generate older face images given baseline ones. There are several similarities between face and brain ageing generation: first, they both aim to generate visual predictions of given current images; second, *subject identity* needs to be maintained for both tasks, *i.e.* the synthetic images should be from the same subject as the inputs; third, the synthetically aged images need to be realistic. As such, approaches for face ageing synthesis have the potential to be applied in the field of brain ageing. For example, the method of [193] has been adopted by the authors of [191, 192] for brain ageing synthesis. In the original work [193], the authors used one-hot vectors to represent different age spans and concatenated the age vector to the latent vectors produced by the encoder. Similarly, a few approaches [199, 200, 201] adopted the conditional GAN framework to generate aged face images. However, these approaches encoded age with one-hot vectors for different age spans, *e.g.* from 40 to 50. The one-hot encoding of age inherently treated face ageing synthesis as a multi-class synthesis problem with a class representing a specific age span and thus ignored the inherent ordering of age.

3.4 Evaluation metrics

One particular challenge for image synthesis is how to evaluate if a synthetic image is good or not. For instance, the original GAN [107] generates an image from a latent vector, and thus there are no ground truth target images to evaluate the synthetic results. Furthermore, due to the adversarial training, the objective functions of GANs could not be used to evaluate its performance. Evaluation measures have surfaced with the emergence of new generative models. Some measures emphasized *qualitative* ways such as visual comparisons and human tests. An example of qualitative evaluation is to let human observers judge if synthetic images are realistic or not, and successfully fooling a person implies the satisfactory quality of generated images. However, this type of measure may not reflect the variety of synthetic images and could be circumvented by mode collapse, *i.e.* generated images are of a small selection. Other measures focused on evaluating the performance of generative models in an objective way. Quantitative measures normally extract features from generated images and

real images using pre-trained deep neural networks and then calculate numeric metrics on the extracted features. Common quantitative metrics include Inception Score (IS) [202], Fréchet Inception Distance (FID) [203], Modified Inception Score (m-IS) [133], AM score [204], *etc.* In general, most of these quantitative metrics focus on two aspects: *fidelity*, *i.e.* how realistic the synthetic image looks like, and *versatility*, *i.e.* how diverse the synthetic images are. Readers are referred to a review paper of evaluation metrics of GANs for more details [205].

Ground-truth paired data are the golden standard if available. When ground-truth paired data are not available, the evaluation becomes challenging. Below we will introduce quantitative metrics that are commonly used to measure image quality.

Let us define 2D images $x, \hat{x} \in \mathcal{X}$, where x is a real image and $\hat{x}$ is a synthetic image from some models, and $\mathcal{X} \subset \mathcal{R}^{H \times W}$ with H and W being the height and width of the image, respectively. Then we can define these metrics as follows.

Mean Squared Error (MSE) is the mean squared difference between two images and is defined as:

$$MSE(x, \hat{x}) = \frac{1}{H \times W} \sum_{h \in H} \sum_{w \in W} [\hat{x}(h, w) - x(h, w)]^2. \tag{3.22}$$

Mean Absolute Error (MAE) is the mean absolute difference and is defined as:

$$MAE(x, \hat{x}) = \frac{1}{H \times W} \sum_{h \in H} \sum_{w \in W} |\hat{x}(h, w) - x(h, w)|. \tag{3.23}$$

Structural Similarity Index (SSIM) [206] measures the similarity between two images and also reflects image quality. Denote μ_x and σ_x as the mean and variance of real image x, $\mu_{\hat{x}}$ and $\sigma_{\hat{x}}$ as the mean and variance of the synthetic image $\hat{x}$, and $\sigma_{\hat{x}x}$ as the covariance between x and $\hat{x}$. SSIM is given by:

$$SSIM(x, \hat{x}) = \frac{(2\mu_x \mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)}, \tag{3.24}$$

where c_1 and c_2 are two variables that stabilise the division with weak denominator.

Peak Signal to Noise Ratio (PSNR) [207] measures the ratio between the maximum possible power of a signal and the power of corrupting noise. PSNR also reflects image quality and is defined as:

$$PSNR(x, \hat{x}) = 10 \cdot \log_{10} \left(\frac{MAX_x^2}{MSE(x, \hat{x})} \right), \tag{3.25}$$

where MAX_x is the maximum pixel value of the image.

Mutual Information (MI) [208] stems from information theory and is a measure of of the mutual dependence between two signals. It was first proposed in [209] to measure registration for multi-modality images assuming that regions of similar tissue (and similar gray values) in one image should correspond to regions in the other image that also consist of similar gray values (though the values may be different from those of the first image). MI is measured by:

$$I(x, \hat{x}) = H(x) + H(\hat{x}) - H(x, \hat{x}), \tag{3.26}$$

where $H(x)$ and $H(\hat{x})$ are the entropy of images x and $\hat{x}$, respectively, and $H(x, \hat{x})$ is the joint entropy of x and $\hat{x}$.

Note SSIM and MI both measure the matching of two images. Their difference is that SSIM focuses more on the perceived similarity between two images and thus requires these two images to share the same modality, while MI does not require pixel values to be the same in the two images and is a measure of how well you can predict pixels in the second image, given pixel values in the first image. In this thesis, we use SSIM to measure the similarity between two images, as we only focus on one modality in a single task. However, if we want to measure the alignment or similarity between images of different modalities, e.g. a T1 and a T2 MRI image, we should use MI as the metric.

For some specific image synthesis tasks such as pseudo healthy synthesis, evaluation needs to consider more than *fidelity* and *versatility*. For instance, in the context of pseudo healthy synthesis, when translating a medical image from pathological to healthy domain, despite fidelity we also care about if the translated image belongs to the same subject as the original image. For proposed metrics specific for our tasks please refer to Section 4.4.4 and

Section 5.4, respectively.

3.5 Summary

This chapter introduced and discussed the technical background of this thesis. We first briefly introduced machine learning and motivated image synthesis. Then we gave an overview of deep generative methods focusing on image synthesis, Generative Adversarial Networks (GANs). Moreover, we introduced the concepts of pseudo healthy synthesis and brain ageing synthesis and reviewed recent progress in these fields. Finally, we introduced the evaluation metrics for image synthesis tasks, particularly focusing on the evaluation for pseudo healthy and brain ageing synthesis.

Chapter 4
Pseudo Healthy Synthesis[1]

4.1 Introduction

In this chapter, we focus on pseudo healthy synthesis. As we briefly introduced in Chapter 1 and 3, the goal of pseudo healthy synthesis is to generate subject-specific 'healthy' images from pathological ones. By definition, a good pseudo healthy image should both be *healthy* and preserve the subject *identity*, *i.e.* belong to the same subject as the input. The synthesis of such 'healthy' images has many potential applications both in research and clinical practice. For instance, synthetic 'healthy' images can be used for pathological segmentation, *e.g.* ischemic stroke lesion, by comparing the real with the synthetic image [9, 159]. Similarly, these 'healthy' images can be used for detecting which part of the brain is mostly affected by neurodegenerative diseases, *e.g.* in Alzheimer disease, a more challenging task because of the global effect of these diseases [12].

However, devising methods that achieve the above task remains challenging. Methods relying on supervised learning are not readily applicable, as finding both pathological and healthy images of the same subject for training and evaluation is not easy, since a subject cannot be 'healthy' and 'unhealthy' at the same time. Even though the use of longitudinal data could perhaps alleviate this, the time difference between observations would introduce more complexity to the task by adding as a confounder ageing alterations on the images beyond the manifestation of the actual disease.

[1] **This chapter is based on the following papers:**

- **Tian Xia, Agisilaos Chartsias, and Sotirios A. Tsaftaris. "Adversarial Pseudo Healthy Synthesis Needs Pathology Factorization". In International Conference on Medical Imaging with Deep Learning, pp. 512-526. PMLR, 2019.**
- **Tian Xia, Agisilaos Chartsias, Sotirios A. Tsaftaris, "Pseudo-healthy synthesis with pathology disentanglement and adversarial learning", Medical Image Analysis, Volume 64, 2020, 101719, ISSN 1361-8415.**

As introduced in Section 3.3.1, prior to the rise of deep learning, approaches were focused on learning manifolds between 'healthy' and 'diseased' local regions at the patch [9, 157] or even voxel level [158]. However, the extent that these methods could capture global alterations of appearance, due to disease, remained limited. Recently though, the advent of deep learning in medical imaging [210] has led to new approaches to pseudo healthy synthesis. Some recent works [163, 160] for example, scaled up the approach of manifold learning to the image level with convolutional architectures. More recently, adversarial approaches allowed learning mappings between the healthy and pathological image domains [12, 171] .

4.1.1 Motivation for our approach

We follow the same spirit, but differently from previous works our method focuses on disentangling the pathological from the healthy information, as a principled approach to guide the synthetic images to be 'healthy' and preserve subject *'identity'*. Figure 4.1(a) illustrates an example of identity loss. Thus, while our goal is to come up with an image that is healthy looking, we also aim to preserve identity such that the generated image belongs to the same input subject.

We use cycle-consistency [130] to help preserve identity but this introduces the so-called *one-to-many problem* (detailed description in Section 4.3.2), where due to lack of information in the pseudo healthy image we may now lose identity in the reconstructed image (see Figure 4.1(b)). Our approach, by disentangling the information related to disease in a separate segmentation mask, circumvents this and helps enable many-to-many mappings (see Figure 4.1(c)).

4.1.2 Overview for our approach

A simple schematic of our proposed 2D method is shown in Figure 4.2. The proposed network contains three components to achieve our goal during training: the *Generator* (G) transforms a pathological image to a pseudo healthy one; the *Segmentor* (S) segments the pathology in the input image; finally, the *Reconstructor* (R) reconstructs the input pathological

image by combining the 'healthy' image with the segmented mask and closes the cycle. The segmentation path is important to preserve the pathological information, and the reconstruction path involving the cycle-consistency loss contributes to the preservation of the subject identity. Note that during inference we only use the Generator and Segmentor.

The proposed method can be trained in a supervised manner using *paired* pathological im-

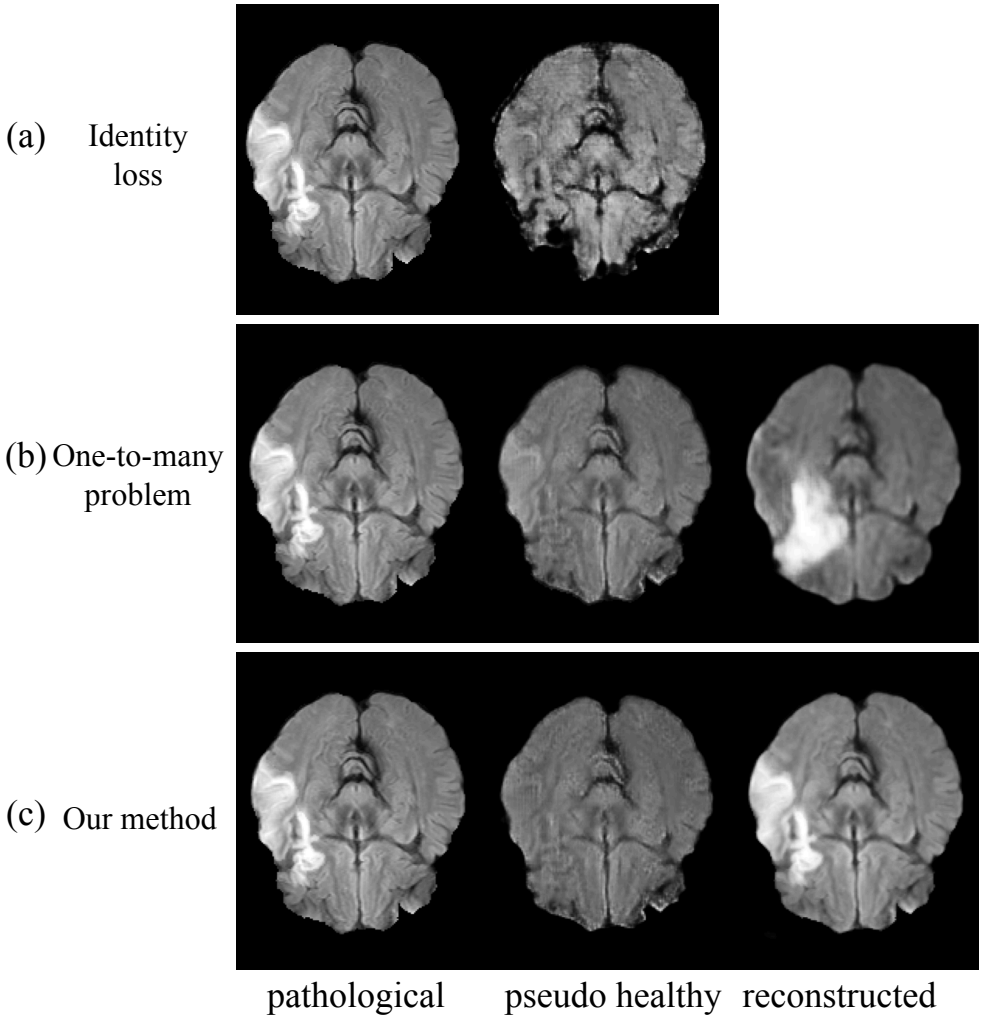


Figure 4.1: The challenge of preserving identity. (a) shows an example of *identity* loss in the generated 'healthy' image. (b) shows a failure example of *one-to-many problem* (described in Section 4.3.2). (c) shows an example obtained by our method which preserves *identity* well. From left to right are the pathological image, pseudo healthy image and the reconstructed image (if any), respectively. The example is taken from the ISLES dataset.

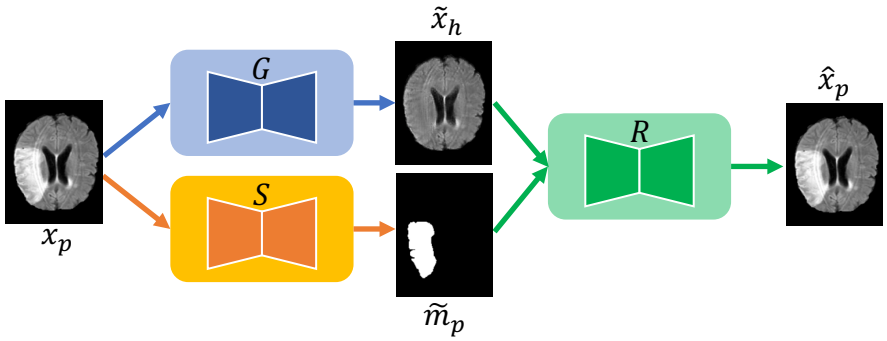


Figure 4.2: Schematic of our approach. A pseudo healthy image $\tilde{x}_h$ is generated from the input pathological image x_p by the *Generator* (G); a pathological mask $\tilde{m}_p$ is segmented from x_p by the *Segmentor* (S); finally a reconstructed image $\hat{x}_p$ is reconstructed from $\tilde{x}_h$ and $\tilde{m}_p$ by the Reconstructor (R).

ages and masks. However, since manually annotating pathology can be time-consuming and requires medical expertise, we also consider an *unpaired* setting, where such pairs of images and masks are not available. Overall, our method is trained with several losses including a cycle-consistency loss [130], but we use a modified second cycle where we enforce healthy-to-healthy image translation to help preserve the identity.

4.1.3 Contributions

The main contributions of this chapter are the following:

- We propose a method for pseudo healthy synthesis by disentangling anatomical and pathological information, with the use of supervised and unsupervised (adversarial) costs.
- Our method can be trained in two settings: *paired* in which pairs of pathological images and masks are available, and *unpaired* in which there are no corresponding segmentations for the input images.
- We introduce quantitative metrics and subjective studies to evaluate the 'healthiness' and 'identity' of the synthetic results, and present extensive experiments comparing

with four different methods (baselines and recent models form the literature), as well as ablation studies, on different MRI modalities.

- We observe that our method may have the capacity of correcting brain deformations caused by high grade glioma, and propose a metric to assess this deformation correction.

- We introduce a subjective study where human raters evaluate the quality of created images.

The rest of this chapter is organised as follows: Section 4.2 reviews the literature related to pseudo healthy synthesis. Section 4.3 presents our proposed method. Section 4.4 describes the experimental setup and Section 4.5 presents the results and discussion. Finally, Section 4.6 concludes the chapter.

4.2 Related work

The concept of medical image synthesis is defined by [211] as '*the generation of visually realistic and quantitatively accurate images*', and the corresponding task has attracted significant attention recently. Here we briefly summarise literature related to pseudo healthy synthesis (*refer to Section 3.3.1 for a more detailed review*), and discuss the differences between our method and these approaches.

4.2.1 Non-deep learning methods

Early methods learned local manifolds at the patch or pixel level [9, 157]. Patches were used together with dictionary learning to learn a linear mapping of source (pathological) and target (healthy) patches. Then, pseudo healthy synthesis can be performed by searching for the closest patches within the dictionary and propagating the corresponding healthy patches to the synthetic 'healthy' image. However, these methods heavily rely on the variation and size of the learned dictionaries. When input pathological patches are not similar to the training

patches, these methods may not find suitable healthy patches to generate the 'healthy' image. Furthermore, these methods are limited by the linear approximation of the dictionary decomposition.

4.2.2 Autoencoder methods

Aiming to scale up the receptive field of these methods and to permit more complex non-linear mappings, deep learning methods were employed first by learning compact manifolds in latent spaces to represent healthy data employing autoencoders [163, 164, 165, 161, 160]. These approaches assume that when abnormal images are given to a neural network trained with healthy data, they are transformed (via the reconstruction function of the autoencoder) to images within the normal (healthy) distribution. Usually non-healthy data are not used in training and guarantees that the synthetic images will maintain subject identity and be indeed within the manifold of the healthy distribution are thus not given. Furthermore, recently the correctness of modelling an input (normal) distribution to detect abnormal, out-of-distribution data has been questioned [212].

4.2.3 Generative models

To involve abnormal data, Generative Adversarial Network (GAN) [107] and its variants [213, 130] can be used. For instance, in [6] and [12] an ℓ_1 regularization loss and an adversarial loss were used to help preserve subject identity. But they either required paired data [6] or apply strong regularization that could conflict with the synthesis process [12] . Another approach to help preserve identity in the unpaired setting is the cycle-consistency loss of CycleGAN [130]. CycleGAN has been adopted for pseudo healthy synthesis of glioblastoma brain images [168, 169, 170] and for liver tumours [171]. However, when one domain contains less information than the other, CycleGAN faces the *one-to-many problem* (described in Section4.3.2, which affects the quality of synthetic images, as mentioned in Section 4.1.1 and highlighted in Figure 4.1 (b). In order to alleviate this problem, pathology was provided as residual and treated tumour as an additive factor in [169, 170]. However, their focus is on segmentation, instead of the quality of synthetic images. Our approach differs from these

methods by treating pathology as a complex factor that can affect the whole brain. In addition, part of the training process involves the Cycle H-H, detailed in Section 4.3.5, to help synthesis.

4.2.4 Our approach

Our approach aims to address the above shortcomings. Similar to CycleGAN, our approach uses cycle-consistency losses to encourage identity preservation, however it also addresses the *one-to-many problem* by disentangling images in pathological and anatomical factors. Thus, we aim to control both processes. In addition, in our effort to demonstrate the capabilities of adversarial approaches, we use as healthy domain images from a different unrelated dataset. This helps correct deformations caused by tumour masses. Finally, as we also noted in Section 4.1, we directly evaluate images explicitly with new metrics, as well as with an observer study, rather than implicitly evaluating quality with performance in downstream tasks.

4.3 Methodology

4.3.1 Problem overview and notation

We denote a pathological image as x_{p_i}, i indicating a subject. x_{p_i} belongs to the pathological distribution, $x_{p_i} \sim \mathcal{P}$. The goal is to generate a pseudo healthy image $\tilde{x}_{h_i}$ for the pathological image x_{p_i}, such that $\tilde{x}_{h_i}$ lies in the distribution of healthy images, $\tilde{x}_{h_i} \sim \mathcal{H}$. We also want the generated image $\tilde{x}_{h_i}$ to maintain the identity of subject i. Therefore, pseudo healthy synthesis can be formulated as two major objectives: *remove* the disease of pathological images, and *maintain* the identity and realism. For ease and unless explicitly stated, in the rest of the chapter, we omit the subscript index i, and directly use x_p and x_h to represent samples from $\mathcal{P}$ and $\mathcal{H}$ distributions, respectively.

4.3.2 The one-to-many problem: for pathology disentanglement

The transformation of a pathological image x_p to its healthy version $\tilde{x}_h$ means that $\tilde{x}_h$ does not have the information of pathology present in the image. The question that arises is then: *How can CycleGAN reconstruct x_p from $\tilde{x}_h$ when this pathology information is lost?* There could be many x_p with disease appearing in different locations that correspond to the same $\tilde{x}_h$. Given this information loss from one domain to the other, CycleGAN has to either hide information within the domain data [172] and/or somehow within the extra capacity of the network to 'permit' it to invent the missing information. An example failure case can be seen in Figure 4.1 (b). We observe that the location and shape of the ischemic lesion is different between the original and reconstructed image. This is because the pseudo healthy image does not contain, anymore, lesion information to guide the reconstruction of the input image.

Recent papers [214, 215, 216] have shown that auxiliary information can be provided in the form of a style or modality specific code (a vector) to guide the translation and permit now a well-posed one-to-one mapping. Our work follows a similar idea and considers the auxiliary information to be spatial, and specifically stores the location and shape of the pathology in the form of a segmentation map. This then overcomes the one-to-many problem, and prevents the decoder from storing disease related features in the weights and the encoder from the need to encode pathology information in the pseudo healthy image.

4.3.3 Proposed approach

An overview of our approach including the training losses is illustrated in Figure 4.3. The proposed method contains three components, the architectures of which are shown in Figure 4.4: the *Generator*, the *Segmentor* (S) and the *Reconstructor* (R). The Generator and the Segmentor comprise the pseudo healthy part of our approach, and disentangle a diseased image into its two components, the corresponding pseudo healthy image and the segmentation mask.

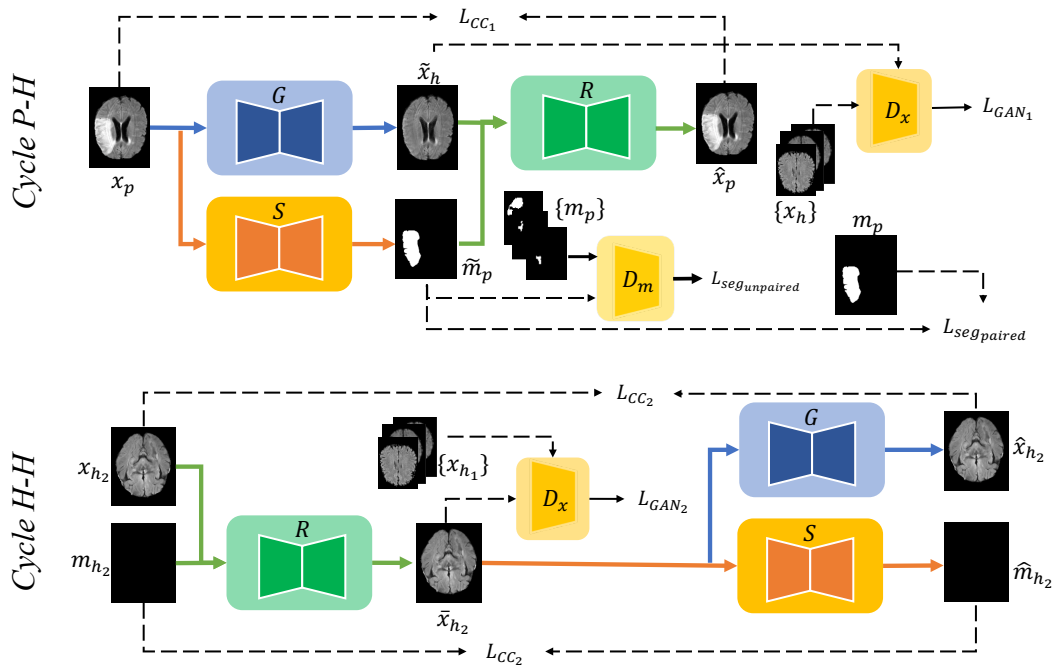


Figure 4.3: Training the proposed method. In *Cycle P-H*, a pathological image x_p is firstly disentangled into a corresponding pseudo healthy image $\tilde{x}_h$ and a pathology segmentation $\tilde{m}_p$. Synthesis is performed by the generator network G and the segmentation by the segmentor S. The pseudo healthy image and the segmentation are further combined in the reconstructor network R to reconstruct the pathological image $\hat{x}_p$. In Cycle H-H, a healthy image x_h and its corresponding pathology map (a black mask) m_h are put to the input of the reconstructor R to get a fake 'healthy' image, denoted as $\bar{x}_h$ to differ from the pseudo healthy image $\tilde{x}_h$ in *Cycle P-H*. This 'healthy' image $\bar{x}_h$ is then provided to G and S to reconstruct the input image and mask, respectively.

4.3.3.1 Generator

The Generator transforms diseased to pseudo healthy images. The Generator architecture has long skip connections between downsampling and upsampling blocks. This helps better preserve details of the input images and results in sharper outputs. The detailed architecture of the Generator is shown in Figure 4.4.

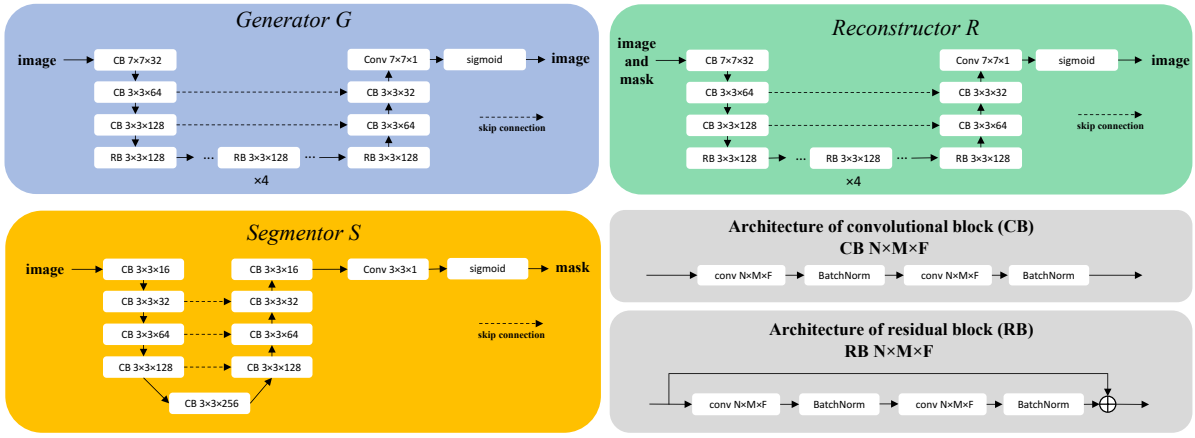


Figure 4.4: Detailed architectures of three main components in our method. The *Generator G* and *Reconstructor R* are modified residual networks [13] with long skip connections between up- and down-sampling blocks. The difference between the Generator and the Reconstructor is that the first takes a one-channel input (image), whereas the second takes a two-channel input (image and mask). The Segmentor is a U-net [14] with long skip connections. All convolutional layers use *LeakyReLU* as activation function, except for the last layers which use *sigmoid*.

4.3.3.2 Segmentor

The Segmentor predicts a binary disease segmentation map.[2] This map helps localise and delineate disease in the reconstructed image. The Segmentor follows a U-net [14] architecture, shown in Figure 4.4.

4.3.3.3 Reconstructor

The Reconstructor takes a pseudo healthy image and a corresponding segmentation mask of the disease, concatenates them in a two-channel image, and reconstructs the input, pathological, image. The architecture of the Reconstructor is the same as the one of the Generator, except that Generator takes one-channel input but Reconstructor takes a two-channel input. Image reconstruction is key for our method since it helps preserve subject identity.

[2]We also investigated using a single neural network with shared layers and two outputs to perform this decomposition, but found that using two separate networks enables more stable training. This architectural choice is in line with other disentanglement methods [217, 218].

4.3.3.4 Discriminators

Our method involves two discriminators that are used in adversarial training. One is the discriminator for pseudo healthy images (denoted as D_x) which encourages generation of realistic pseudo healthy images. The other is used to help learn a manifold for the pathology mask (denoted as D_m) which is used to train the Segmentor when paired pathological images and masks are not available (more details in Section 4.3.5). The architecture of both discriminators follow the design used by [12]. The adversarial training is performed with a Wasserstein loss with gradient penalty [112].

4.3.4 Model training

Inspired by [130], we involve two cycles to train our model, which are shown in Figure 4.3. The first cycle is *Cycle P-H*, where we perform pseudo healthy synthesis. The Generator G first takes a pathological image x_p as input, and produces a pseudo healthy image: $\tilde{x}_h = G(x_p)$. Similarly, the Segmentor S takes x_p as input and outputs a mask $\tilde{m}_p$ indicating where the pathology is: $\tilde{m}_p = S(x_p)$. The Reconstructor R then takes both $\tilde{x}_h$ and $\tilde{m}_p$ as input and generates a reconstruction of the input image: $\tilde{x}_p = R(\tilde{x}_h, \tilde{m}_p)$.

The second cycle is *Cycle H-H* which is designed to stabilise training, help preserve input identity, and further encourage disentanglement of disease from the pseudo healthy image. The Reconstructor first takes as input a healthy image x_h and a 'healthy' mask m_h , *i.e.* an image of all zeros, and produces a fake healthy image: $\bar{x}_h = R(x_h, m_h)$. This fake healthy image $\bar{x}_h$ is then passed as input to the Generator G, $\hat{x}_h = G(\bar{x}_h)$, and Segmentor to reconstruct the input healthy image and mask, $\hat{m}_h = S(\bar{x}_h)$, respectively.

The design of Cycle H-H is due to several reasons. First, we want to ensure that the Reconstructor does not invent pathology when given a healthy mask as input. Second, we encourage the Generator to better preserve identity, *i.e.* when the input to G is a 'healthy' image, the output should be the same 'healthy' image. Similarly, when given a 'healthy' image, the Segmentor should not detect any pathology. When the predicted output is not a black map, it means that either the Reconstructor is not trained well, *i.e.* it creates pathology-like artefacts, or the Segmentor is not trained well, *i.e.* it finds non-existing pathology. In this case, the

Reconstructor and Segmentor are penalised. This in turn also encourages the Segmentor not to hide information useful for reconstruction, and thus any anatomical information is only contained in the pseudo healthy image.[3]

4.3.5 Paired and unpaired settings

There are two settings of training the Segmentor (S) considering the availability of ground-truth pathology labels.

In the first, termed *paired* setting, we have paired pathological images and ground-truth masks. In this setting, we train the Segmentor directly using the ground-truth pathology masks with a differential analogue of the Dice segmentation loss.

In the second, termed *unpaired* setting, we do not have pairs of pathological images and masks. In this setting, since supervised training is not feasible, we involve a *Mask Discriminator* termed as D_m that distinguishes segmented masks from real pathology masks, and thus learns a prior on the pathology shape. The Segmentor is then trained adversarially against this Mask Discriminator. The real pathology masks used for training are ground-truth pathology masks chosen randomly from other subjects. The losses are described mathematically for each setting in Section 4.3.6.3.

4.3.6 Losses

The training losses can be divided into three categories, *adversarial losses*, *cycle-consistency losses* and *segmentation losses*, the details of which are described below.

[3]We note here that we could also have considered a cycle where we could take a pseudo healthy image and pass it through the segmentor and penalise if any disease pixels are detected. We found that this is less stable: either the segmentor could have thrown a false positive or the generator made an error. We found the design of the current Cycle H-H more robust and our experiments show that the pseudo healthy images rarely contain detectable, by a judge segmentor, disease pixels.

4.3.6.1 Adversarial losses for images

The synthesis of pseudo healthy image $\tilde{x_h}$ ($\tilde{x_h} = G(x_p)$) in *Cycle P-H* is trained using the Wasserstein loss with gradient penalty [112]:

$$L_{GAN_1} = \max_{D_x} \min_{G} \mathbb{E}_{x_p \sim \mathcal{P},\, x_h \sim \mathcal{H}}[D_x(x_h) - D_x(G(x_p)) \\ + \lambda_{GP}(\|\nabla_{\dot{x}_h}(\dot{x}_h)\|_2 - 1)^2], \tag{4.1}$$

where x_p is a pathological image, $G(x_p)$ is its corresponding pseudo healthy image, x_h is a healthy image, D_x is the discriminator to separate real and fake samples, and $\dot{x}_h$ is the average sample defined by $\dot{x}_h = \epsilon\, x_h + (1-\epsilon)\, G(x_p)$, $\epsilon \sim U[0,1]$. The first two terms measure the Wasserstein distance between real healthy and synthetic healthy images; the last term is the gradient penalty loss involved to stabilise training. As in [112] and [12], we set $\lambda_{GP} = 10$.

Similarly, we have L_{GAN_2} for the fake 'healthy' image $\bar{x}_h$ ($\bar{x}_h = R(x_h, m_h)$) in Cycle H-H:

$$L_{GAN_2} = \max_{D_x} \min_{R} \mathbb{E}_{x_{h_1} \sim \mathcal{H},\, x_{h_2} \sim \mathcal{H},\, m_{h_2} \sim \mathcal{H}_m}[D_x(x_{h_1}) \\ - D_x(R(x_{h_2}, m_{h_2})) + \lambda_{GP}(\|\nabla_{\dot{x}_h}(\dot{x}_h)\|_2 - 1)^2], \tag{4.2}$$

where x_{h_1} and x_{h_2} are two different healthy images drawn from the healthy image distribution $\mathcal{H}$, m_{h_2} is the corresponding pathology mask of x_{h_2}, *i.e.* a black mask, $R(x_{h_2}, m_{h_2})$ is the fake 'healthy' image reconstructed with x_{h_2}, and $\dot{x}_h$ is defined as $\dot{x}_h = \epsilon\, x_{h_1} + (1-\epsilon)\, R(x_{h_2}, m_{h_2})$, $\epsilon \sim U[0,1]$.

4.3.6.2 Cycle-consistency losses

We involve cycle-consistency losses to help preserve the subject identity of the input images. For *Cycle P-H*, we have:

$$L_{CC_1} = \min_{G,R,S} \mathbb{E}_{x_p \sim \mathcal{P}}[\|R(G(x_p), S(x_p)) - x_p\|_1], \tag{4.3}$$

where x_p is a pathological image, $G(x_p)$ is the pseudo healthy image produced by Generator, $S(x_p)$ is the segmented pathology mask by Segmentor, $R(G(x_p), S(x_p))$ is the reconstructed

pathological image by Reconstructor given $G(x_p)$ and $S(x_p)$. Similarly with [130], we use ℓ_1 loss rather than ℓ_2, to reduce the amount of blurring.

Similarly, for Cycle H-H, we have:

$$L_{CC_2} = \min_{G,R,S} \mathbb{E}_{x_{h_2}\sim\mathcal{H},\, m_{h_2}\sim\mathcal{H}_m}[\|G(R(x_{h_2}, m_{h_2})) - x_{h_2}\|_1 + \|S(R(x_{h_2}, m_{h_2})) - m_{h_2}\|_1], \tag{4.4}$$

where x_{h_2} and m_{h_2} are a healthy image and the corresponding mask, respectively, $R(x_{h_2}, m_{h_2})$ is the fake 'healthy' image obtained by Reconstructor given a healthy image x_{h_2} and a healthy mask m_{h_2} as input, $G(R(x_{h_2}, m_{h_2}))$ is the reconstructed image by Generator given $R(x_h, m_{h_2})$, and $S(R(x_{h_2}, m_{h_2}))$ is the segmented mask that corresponds to $R(x_{h_2}, m_{h_2})$. Here we use ℓ_1 loss for the reconstructed mask instead of the Dice loss as it is not well defined when the target masks are all black.

4.3.6.3 Segmentation losses

As described in Section 4.3.5, there are two training settings for the Segmentor. For the paired setting where we have access to paired pathological image and masks, we use a supervised loss to train the Segmentor:

$$L_{seg_{paired}} = \min_{S} \mathbb{E}_{x_p\sim\mathcal{P},\, m_p\sim\mathcal{P}_m}[Dice(m_p, S(x_p))], \tag{4.5}$$

where x_p and m_p are paired pathological images and masks, $S(x_p)$ is the predicted mask by *Segmentor S*, and $Dice(.)$ represent the dice coefficient loss [219].

In the unpaired setting, there are no paired images and masks, and we use an adversarial loss to train the Segmentor:

$$L_{seg_{unpaired}} = \max_{D_m} \min_{S} \mathbb{E}_{x_{p_1}\sim\mathcal{P},\, m_{p_2}\sim\mathcal{P}_m}[D_m(S(x_{p_1})) - D_m(m_{p_2}) + \lambda_{GP}(\|\nabla_{\bar{m}_p} D(\bar{m}_p)\|_2 - 1)^2], \tag{4.6}$$

where x_{p_1} is a pathological image, m_{p_2} is a pathological mask randomly drawn from subjects other than x_{p_1}, D_m is the discriminator to classify between the segmented mask $S(x_{p_1})$ and the randomly chosen mask m_{p_2}, and $\bar{m}_p$ is the average sample defined by $\bar{m}_p = \epsilon\, m_{p_2} + (1 - \epsilon)\, S(x_{p_1})$, $\epsilon \sim U[0, 1]$.

4.4 Experimental setup

4.4.1 Data and pre-processing

Data: In this work we use 2D slices from three datasets: ISLES, BraTS and CamCAN, which are described in Section 2.5. Here we detail the modalities and number of subjects of each dataset that are used for our experiments.

- *Ischemic Stroke Lesion Segmentation challenge 2015* (ISLES) contains 28 volumes. All volumes have lesion segmentation annotated by experts. We use T2 and FLAIR modality for our experiment.
- *Multimodal Brain Tumor Segmentation Challenge 2018* (BraTS) [4] dataset contains high and low grade glioma cases. In this work we select 150 volumes which contain high grade glioma/glioblastoma (HGG). The 'healthy' slices in BraTS may not be really healthy, since the glioblastoma may affect areas of brain where it is not present [4], for an example see Figure 4.7. We therefore involve Cam-CAN dataset as a healthy dataset, as described below.
- *Cambridge Centre for Ageing and Neuroscience* (Cam-CAN) [220] dataset contains normal volumes from 17 to 85 years old. We randomly selected 76 volumes for our experiment. We chose to involve this dataset as 'healthy' data when performing pseudo healthy synthesis to avoid the possible deformations of brain tissues in BraTS images. Since Cam-CAN only contains T1 and T2 modalities, we also use T1 and T2 from BraTS.

Pre-processing: Initially, we skull-stripped the Cam-CAN volumes using FSL-BET [90]. We then linearly registered the Cam-CAN and BraTS volumes to MNI 152 space using

FSL-FLIRT [221]. We normalised the volumes of all datasets by clipping the intensities to $[0, V_{99.5}]$, where $V_{99.5}$ is the 99.5% largest intensity value in the corresponding volume, and rescaled to the range $[0, 1]$. We then selected the middle 60 2D axial slices from each volume, and cropped each slice to the size $[208, 160]$. Note the training and testing are performed on these selected 2D slices, instead of all available 2D slices, because some 2D slices do not contain any brain parts or only contain a small portion and thus do not provide much information. For ISLES, we label a slice as 'healthy' if its corresponding lesion map is black, otherwise as 'pathological'. We label all slices from Cam-CAN as 'healthy', and label a slice from BraTS as 'pathological' if its corresponding pathology annotation is not a black mask, *i.e.* the glioblastoma is present in this slice.

Histogram check: To check the feasibility of training on the two datasets, we first checked the histogram similarity between BraTS and Cam-CAN. Specifically, we normalised each histogram to a probability density distribution (PDF), and computed the Jensen–Shannon (JS) divergence [222] between the PDFs of the two datasets. We calculated a JS divergence of 0.009 between BraTS 'healthy' slices (slices with no segmentations) and Cam-CAN slices, 0.011 between BraTS 'healthy' and BraTS 'pathological' slices, and 0.015 between BraTS 'pathological' and Cam-CAN slices. This implies that after pre-processing, the difference between histograms of Cam-CAN and BraTS is minimal, and thus intensities distributions of them are close.

4.4.2 Baselines and methods for comparison

We compare our method with the following four approaches. These methods are introduced in Section 3.2.1 and 3.3.1. Here we briefly describe how they are comparable to our method :

1. **Conditional GAN:** We first consider a baseline that uses adversarial training and a simple conditional approach of [122]. This is a GAN in which the output is conditioned on the input image and does not use segmentation masks. This baseline uses a generator and a discriminator with the same architectures as our method for appropriate comparison.

2. **CycleGAN:** Another baseline we compare with is the CycleGAN [130], where there are two translation cycles: one is *P* to *H* to *P*, and the other is *H* to *P* to *H* ('*P*' refers to the pathological and '*H*' refers to the healthy domain). We do not use segmentation masks. The generators and discriminators of CycleGAN also share the same architecture as our proposed method.

3. **AAE:** We implement and compare with a recent method that aims to address a similar problem [160]. We trained an adversarial autoencoder (AAE) only on healthy images and performed pseudo healthy synthesis with the trained model. This approach does not use segmentation masks and data with pathology.

4. **vaGAN:** We compare with [12], another recent method for pseudo healthy synthesis, using the official implementation[4] but modified for 2D slices. This method produces residual maps, which are then added to the input images to produce the resulting pseudo healthy images. An ℓ_2 loss on the produced maps acts as a regulariser. This approach does not use segmentation masks.

4.4.3 Training details

In the paired setting, the overall loss is:

$$\begin{aligned} L_{paired} &= \lambda_1 L_{GAN_1} + \lambda_2 L_{GAN_2} \\ &+ \lambda_3 L_{CC_1} + \lambda_4 L_{CC_2} + \lambda_5 L_{seg_{paired}}, \end{aligned} \tag{4.7}$$

where the λ parameters are set to: $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 20$, $\lambda_4 = 10$ and $\lambda_5 = 10$.

In the unpaired setting, the loss is:

$$\begin{aligned} L_{unpaired} &= \lambda_1 L_{GAN_1} + \lambda_2 L_{GAN_2} \\ &+ \lambda_3 L_{CC_1} + \lambda_4 L_{CC_2} + \lambda_5 L_{seg_{unpaired}}, \end{aligned} \tag{4.8}$$

where λ_1, λ_2, λ_3 and λ_4 are set as above, while λ_5 is set to 1. The values of the λ parameters are set experimentally as follows. The λ for Cycle P-H are double the λ for Cycle H-H, *i.e.*

[4] `https://github.com/baumgach/vagan-code`

$\lambda_1 = 2\lambda_2$ and $\lambda_3 = 2\lambda_4$, since our focus is on pseudo healthy synthesis. Furthermore, the λ for L_{CC} is 10 times larger than the one for L_{GAN} to balance the loss values, *i.e.* $\lambda_3 = 10\lambda_1$ and $\lambda_4 = 10\lambda_2$. Finally, λ_5 in paired setting is set to 10 to encourage an accurate segmentation, since segmentation is a challenging task. The λ values for the unpaired setting are set similarly, except λ_5 that is set to 1, since this is a GAN loss, and a balance between the segmentor and mask discriminator losses is sought. The values of λs were set experimentally. In practice, we found the model is sensitive to loss weights and other hyper-parameters, which could be due to the instability of GAN training and the use of multiple losses.

We train all models for 300 epochs. Following [107] and [111], we updated the discriminators and generators in an alternating session. As Wasserstein GAN requires the discriminators to be close to optimal during training, we updated the discriminators for 5 iterations for every generator update. Initially in the first 20 epochs, we update the discriminators for 50 iterations per generator update. We implemented our methods using Keras [223]. We trained using *Adam* optimiser [224] with a learning rate of 0.0001 and β_1 equal to 0.5. We made the implementation publicly available at `https://github.com/xiat0616/pseudohealthy-synthesis`.

The results of Section 4.5 are obtained from a 3-fold cross validation. For ISLES, each split contains 18 volumes for training, 3 volumes for validation and 7 volumes for testing. For BraTS, each split contains 100 volumes for training, 15 for validation and 35 for testing. For Cam-CAN, each split contains 50 volumes for training, 8 for validation and 18 for testing. This is to ensure that the 'pathological' slices from BraTS have similar number as the 'healthy' slices from Cam-CAN. We fine-tuned the architecture of the pre-trained segmentor and classifier based on the validation set.

4.4.4 Evaluation metrics

Since paired healthy and pathological images of the same subjects are difficult to acquire, we do not have ground-truth images to directly evaluate the synthetic outputs.

As we mentioned previously in Section 4.1.3, image quality has been rarely directly evaluated. To address this, we propose two numerical evaluation metrics to assess the 'healthiness'

and 'identity' of synthetic images [225]. Since brain tumours can cause deformations in the brain, we also propose a new metric to evaluate how well the deformations are corrected in BraTS. Furthermore, we perform human evaluation studies on a subset of our experiments regarding these metrics. Below we introduce the quantitative metrics used in this work.

Healthiness (h): To evaluate how 'healthy' the pseudo healthy images are, we measure the size of their segmented pathology as a proxy. To this end, we pre-trained a segmentor to estimate pathology from images. We then used this segmentor as a judge to assess pathology from the pseudo healthy images and checked how large the estimated pathology areas are. Note that for each split we trained a segmentor on the training data and fine-tuned it on the validation set. Formally, *healthiness* is defined as:

$$h = 1 - \frac{\mathbb{E}_{\hat{x}_h \sim \mathcal{H}}[N(f_{pre}(\hat{x}_h))]}{\mathbb{E}_{m_p \sim \mathcal{P}_m}[N(f_{pre}(x_p))]} = 1 - \frac{\mathbb{E}_{x_p \sim \mathcal{P}}[N(f_{pre}(G(x_p)))]}{\mathbb{E}_{m_p \sim \mathcal{P}_m}[N(f_{pre}(x_p))]}, \tag{4.9}$$

where x_p is a pathological image, f_{pre} is the pre-trained segmentor, and $N(.)$ is the number of pixels that are labelled as pathology by f_{pre}. The denominator uses the segmented mask of the pathological image $f_{pre}(x_p)$, instead of the ground truth m_p, to cancel out a potential bias introduced by the pre-trained segmentor. We subtract the term from 1, such that when pathology mask gets smaller, *h* increases.

Identity (iD): This metric represents how well the synthetic images preserve subject identity, *i.e.* how likely they come from the same subjects as the input images. This is achieved by evaluating their structural similarity to the input images outside the pathology regions, using a masked *Multi-Scale Structural Similarity Index* (MS-SSIM)[5] with window width of 11 [206]. Formally, *identity* is defined as:

$$\begin{aligned} iD &= MS\text{-}SSIM[(1-m_p) \odot \tilde{x_h}, (1-m_p) \odot x_p] \\ &= MS\text{-}SSIM[(1-m_p) \odot G(x_p), (1-m_p) \odot x_p], \end{aligned} \tag{4.10}$$

where x_p is a pathological image, m_p is its corresponding pathology mask, and $\odot$ is pixel-by-pixel multiplication.

[5]Due to its use of MS-SSIM this metric also reflects image quality.

Deformation correction (DeC): In some cases (BraTS dataset), a brain may also deform due to the presence of a large cancerous mass. The difficulty is that, to fix the deformation caused by the tumour, we need to not only change the abnormal intensities, but also to make necessary changes to the structure of the brain. This poses a significant challenge for evaluating subject identity. The identity metric above does not measure well whether this tissue has recovered (because it relies on pixel correspondence). Herein we attempt to define a proxy metric that aims to assess whether such correction has taken place.[6]

As Cam-CAN and BraTS were acquired differently, and could potentially have intensity differences, we pre-processed all brain slices using the Canny edge detector in order to remove any intensity bias. An example of a BraTS image and its extracted edge map are shown in Figure 4.5, where we can observe the deformations as pointed out by the red arrows. We then pre-trained a classifier to classify edge maps of BraTS 'healthy' slices, *i.e.* images with no tumour annotation, and Cam-CAN slices. The pre-trained classifiers, achieved an average accuracy of 89.7%, and were used as a judge on pseudo healthy images from BraTS slices. This means that the classifiers were able to discriminate between BraTS 'healthy' edges and Cam-CAN edges mostly relying on the presence of deformations. The output of this classifier is a continuous number between 0 and 1, representing the probability of an image to be deformation-free. DeC in the testing set is then defined as the probability of synthetic images being deformation-free, *i.e.* more Cam-CAN like.

Human evaluation: To highlight the difficulty of defining quantitative metrics, and the overall difficulty of assessing image 'quality' in such synthesis tasks, we introduce an expert evaluation to further assess the above criteria of healthiness, identity and deformation correction on a small subset of the experiments. We purposely did not ask raters to assess overall image quality, as quality can be a combination of factors (which can vary across experts).[7]

We randomly selected 50 slices from BraTS, obtained the pseudo healthy outputs of all comparison methods, and then asked four medical image analysis researchers and a clinical neu-

[6]We note that this is a very hard task and our attempts to use a non-linear registration-based approach where we measured the amount of deformation between different diseased and pseudo healthy images was not met with success because it gave lots of false positives when identity was completely lost.

[7]We also note the difference of our study design compared to the ones commonly encountered in the image-to-image translation community [130] where users are asked to decide if an image is 'real' or 'fake'.

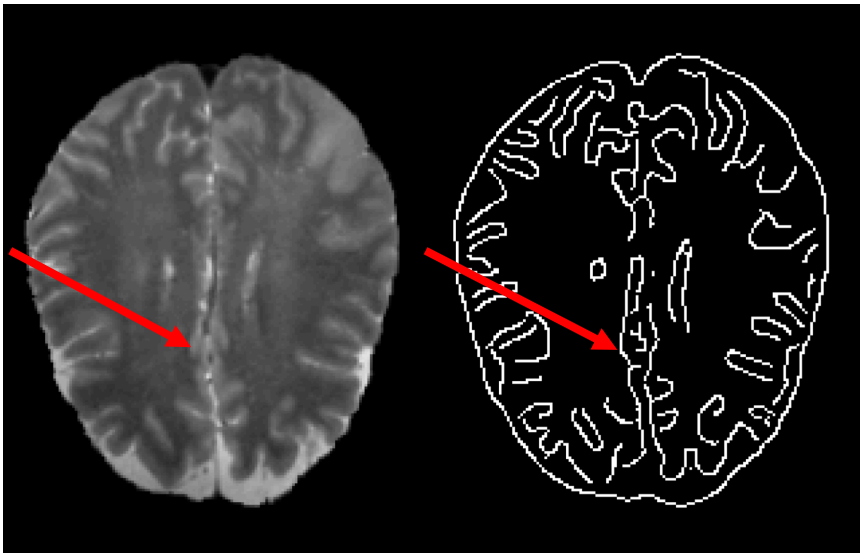


Figure 4.5: An example of BraTS 'healthy' image and its edge map. Observe the deformation in the brain and edge as pointed out by the red arrows. Note that this brain image does not have pathology in its corresponding segmentation map, but the deformation still exists.

rologist to independently score each synthetic image arranged in panels (details below) on each criterion using a binary score. We provided instructions as to what each criterion should reflect. Specifically the definitions were: "Healthiness: assess if the synthetic image appears healthy (1) or not (0)"; "Identity: assess if the synthetic image belongs to the same subject as the original image (1) or not (0)"; "Deformation correction: assess if the deformation caused by a cancerous mass has been corrected in areas outside the mass (1) or not (0)".

Each panel was a montage of: input diseased image; ground truth segmentation mask; pseudo healthy images obtained as outputs of the tested algorithms. The raters were blinded to which algorithm generated each image and image arrangement was randomised (for every panel shown). The raters knew though that the first image was the input to the algorithms.

Overall each rater reviewed 50 panels, each containing 6 images, with a score for 3 metrics, providing a total of 900 scores. Across the four raters 3600 scores were available. We asked raters to limit time spent on a panel to be less than 3 minutes.

Real v.s. fake test: As our approach focuses on image synthesis, we performed a human experiment where we requested raters to tell apart real from synthetic images. Specifically, we

randomly selected 50 pathological slices, and used the methods discussed herein to generate corresponding pseudo healthy images. As a result, we generated 300 images in total. Then, we randomly selected 300 real healthy images, and presented all images in a random order to four researchers who classified them as real or fake. We used a standardised viewing setting (screen size, distance from screen, illumination, monitor brightness) and limited evaluation time to 1 minute per image, and measured 'realness' as the ratio of images labelled 'real'.

4.5 Results and discussion

All results reflect testing sets and we report both averages and standard deviation. We use bold font to denote the best performing method (for each metric) and an asterisk (*) to denote statistical significance compared to the best performing baseline or comparison method (to keep in check multiple comparisons). We use a simple paired t-test to test the null hypothesis that there is no difference between our methods and the best performing baseline, at the significance level of 5%. We found that differences are normally distributed in the quantitative metrics based on the D'Agostino and Pearson's normality test [226, 227]).

4.5.1 Pseudo healthy synthesis for ischemic lesions

Here we perform pseudo healthy synthesis on ISLES dataset, which contains diseased subjects with ischemic lesions. These lesions should not alter the brain's shape distal to the lesion much [20], but rather manifest as hyper-intense regions in T2 and FLAIR modalities. As described in Section 4.4.1, all methods are trained with a 'healthy' set containing images that do not have an annotated lesion mask, and with a 'pathological' set containing the remaining images. The exception is the AAE [160], which requires only 'healthy' images for training. For our method in the unpaired setting, we used approximately 100 masks from 3 subjects for training the mask discriminator. Standard spatial augmentations have been applied to prevent overfitting of the discriminator on the real masks. Note that the baseline and comparison methods do not require pathological masks for training.

We compare our method with the methods of Section 4.4.2 qualitatively and quantitatively.

Numerical results of identity (iD) and healthiness (h), defined in Section 4.4.4, are summarised in Table 4.1, and examples of synthetic images are shown in Figure 4.6.

In Table 4.1 we can see that our method trained in the paired setting achieves the best results, followed by our method trained in the unpaired setting. Both paired and unpaired versions outperform all others. A key reason behind our methods' improved performance is the pathology disentanglement, which enables the accurate reconstruction of the input pathological images without hiding pathology information in the pseudo healthy images. We can also observe from Figure 4.6 that our methods produce sharp and lesion-free images, evidenced also by the superior healthiness values in Table 4.1. The synthetic images also preserve details of the input images, which points that subject identity is preserved along with image quality.

Furthermore, we observe (Table 4.1) that CycleGAN achieves the third best results in terms of *identity*, which showcases the benefit of cycle-consistency loss in preserving subject identity. However, as described in Section 4.3.2, CycleGAN suffers from the *one-to-many problem*, which misleads it to generate artifacts in synthetic images. As a result, the healthiness of CycleGAN is not as good as the ones of vaGAN and Conditional GAN, which do not need to

Method	T2		FLAIR	
	iD	h	iD	h
AAE	$0.63_{0.07}$	$0.71_{0.14}$	$0.66_{0.06}$	$0.81_{0.09}$
vaGAN	$0.72_{0.05}$	$0.77_{0.11}$	$0.75_{0.04}$	$0.85_{0.08}$
Cond. GAN	$0.75_{0.06}$	$0.74_{0.12}$	$0.73_{0.05}$	$0.83_{0.12}$
CycleGAN	$0.82_{0.04}$	$0.76_{0.11}$	$0.83_{0.05}$	$0.81_{0.08}$
Ours (unpaired)	$0.93^{*}_{0.04}$	$0.84^{*}_{0.09}$	$0.87_{0.04}$	$0.88^{*}_{0.06}$
Ours (paired)	$\mathbf{0.97}^{*}_{0.04}$	$\mathbf{0.85}^{*}_{0.08}$	$\mathbf{0.94}^{*}_{0.03}$	$\mathbf{0.89}^{*}_{0.07}$

Table 4.1: Numerical evaluation of our method and baselines on ISLES dataset in terms of *identity* iD and *healthiness* h. For each metric, 1 is the best and 0 is the worst. The best mean values are shown in **bold**. Statistical significant results (5% level) of our methods compared to the best baseline are marked with an asterisk (*).

'hide' pathology information in the pseudo healthy images.

Although vaGAN involves a ℓ_1 loss between the input and synthetic images, we do not see significant improvements over Conditional GAN, where such a regularization loss is not used. In Figure 4.6, we also observe a loss of subject identity in both vaGAN and Conditional GAN. Even though vaGAN produces results that maintain the outline of the brain, these results lack refined details. On the contrary, Conditional GAN changes the outline of the brain but maintains inner details.

In addition, AAE often loses subject identity, and the produced synthetic images may present artifacts within the pathological areas of the input images. This is because there is no explicit loss to force the synthetic images to maintain the subject identity, neither a loss to explicitly ensure that the network learned to transform the pathological area to be 'healthy'.

4.5.2 Pseudo healthy synthesis for brain tumours

Here we apply our method on the BraTS dataset where volumes have high grade glioma. As described in Section 4.5.1, for the case of ischemic lesions we used 'healthy' images from the same dataset. However, as shown in Figure 4.7, BraTS slices with no tumour annotations may still exhibit deformations. As such, training with 'healthy' slices might only adjust the intensities within in the tumour areas, but was not able to fix the deformations caused by tumours. We therefore use a second healthy dataset, Cam-CAN, to extract 2D healthy slices, which we used for model training, after confirming its suitability by comparing its intensity distribution with the one of BraTS (see Section 4.4.1). For our method in the unpaired setting, and to train the mask discriminator, we used approximately 950 masks from 70 subjects that were not part of the training, validation and test sets. Standard spatial augmentations were applied to prevent overfitting of the discriminator on the real masks.

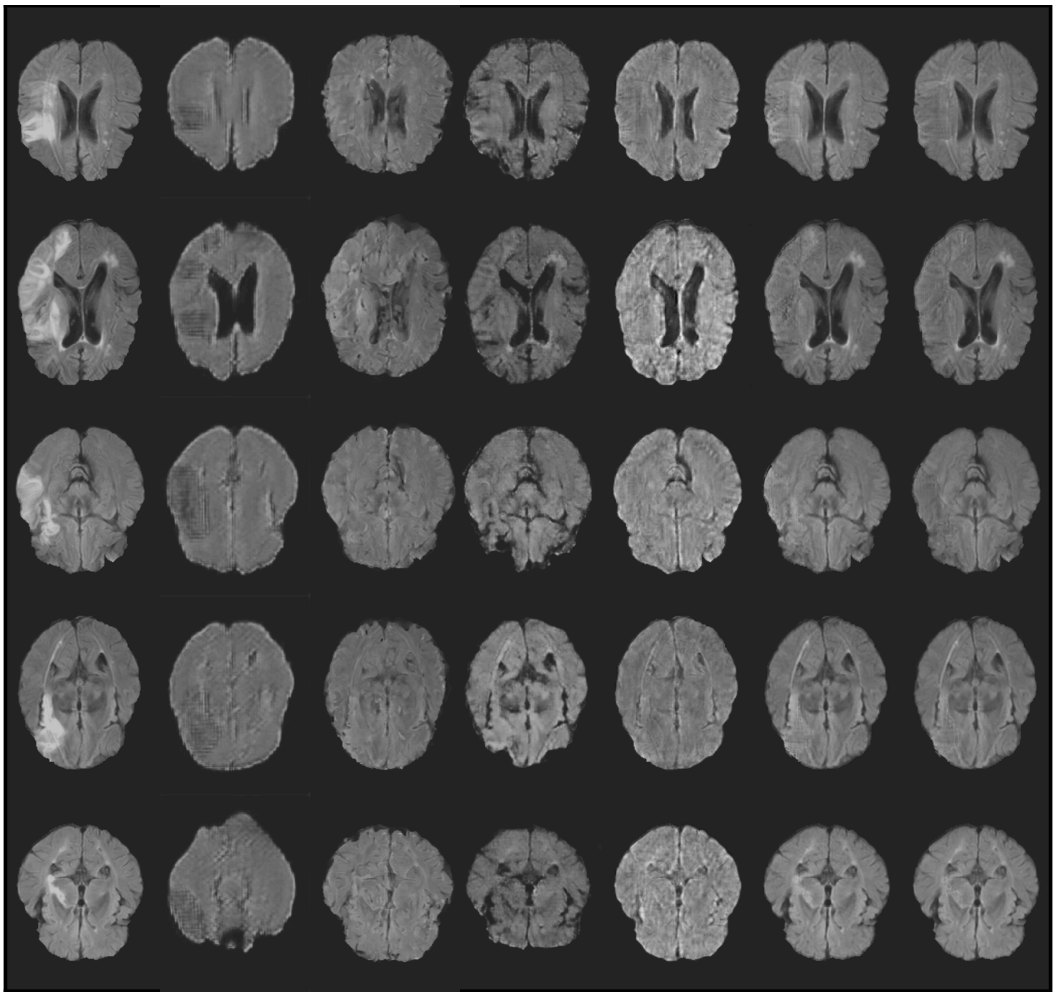


Figure 4.6: Experimental results of five samples (each in every row) for ISLES data. The columns from left to right are the original pathological images, and the synthetic healthy images by *AAE*, *vaGAN*, *Conditional GAN*, *CycleGAN*, and the proposed method in the *unpaired* and *paired* setting, respectively.

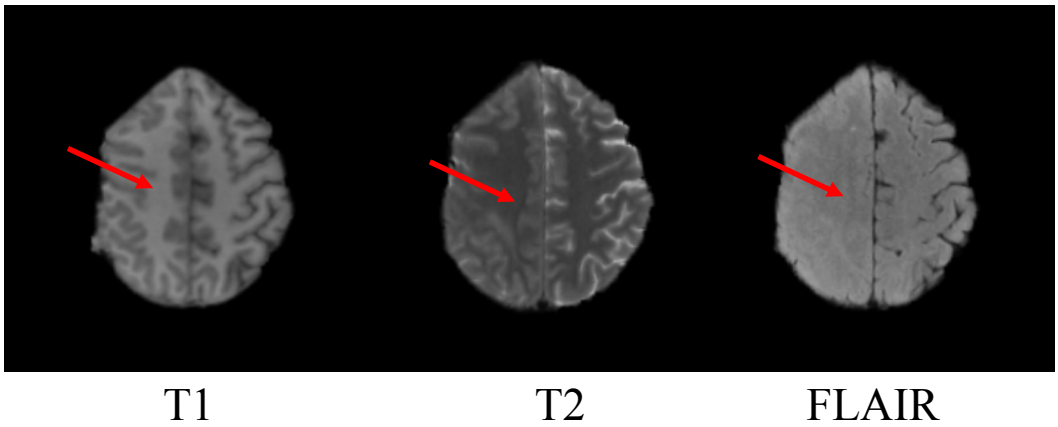


Figure 4.7: An example of BraTS images where glioblastoma is not present, but the brain tissues are still affected by deformations. From left to right are the same slice in T1, T2 and FLAIR modalities, respectively. The red arrows point to the affected areas, *i.e.* the left half of the brain.

Method	T1			T2			T2 (human evaluation)		
	iD	h	DeC	iD	h	DeC	'identity'	'healthiness'	'def. corr.'
AAE	$0.65_{0.12}$	$0.72_{0.16}$	$0.71_{0.05}$	$0.63_{0.12}$	$0.71_{0.13}$	$0.75_{0.04}$	$0.39_{0.34}$	$0.30_{0.32}$	$0.28_{0.31}$
vaGAN	$0.72_{0.11}$	$0.79_{0.12}$	$0.84_{0.06}$	$0.74_{0.10}$	$0.78_{0.09}$	$0.81_{0.05}$	$0.52_{0.34}$	$0.49_{0.33}$	$0.46_{0.39}$
conditional GAN	$0.70_{0.14}$	$0.73_{0.17}$	$0.82_{0.04}$	$0.69_{0.09}$	$0.73_{0.15}$	$0.84_{0.04}$	$0.47_{0.32}$	$0.46_{0.34}$	$0.50_{0.31}$
CycleGAN	$0.82_{0.08}$	$0.80_{0.13}$	$0.71_{0.09}$	$0.81_{0.07}$	$0.77_{0.14}$	$0.73_{0.06}$	$0.56_{0.34}$	$0.53_{0.35}$	$0.30_{0.21}$
Ours (unpaired)	$\mathbf{0.84}_{0.08}$	$0.82_{0.11}$	$\mathbf{0.88}^{*}_{0.11}$	$0.83_{0.06}$	$0.83^{*}_{0.09}$	$0.86^{*}_{0.05}$	$0.65_{0.29}$	$0.67^{*}_{0.27}$	$0.62^{*}_{0.25}$
Ours (paired)	$0.83_{0.06}$	$\mathbf{0.86}^{*}_{0.10}$	$0.85^{*}_{0.10}$	$\mathbf{0.85}^{*}_{0.04}$	$\mathbf{0.84}^{*}_{0.07}$	$\mathbf{0.88}^{*}_{0.04}$	$\mathbf{0.67}^{*}_{0.24}$	$\mathbf{0.69}^{*}_{0.23}$	$\mathbf{0.65}^{*}_{0.25}$

Table 4.2: Results of our methods on BraTS dataset. Here we evaluate three metrics, defined in Section 4.4.4 on T1 and T2 modalities. For each metric, 1 is the best and 0 is the worst. We show also results (last three columns) of a human evaluation on the T2 modality based on criteria as described in Section 4.4.4. The best mean values are shown in **bold**. Statistical significant results (5 % level) of our methods compared to the best baseline are marked with an asterisk (*). 'def. corr.' is a shorthand for 'deformation correction' assessment score from the raters.

Figure 4.8 shows visual comparisons between the methods considered. We observe that our method produces realistic results and preserves details, while other methods are more susceptible to losing subject identity. CycleGAN can better preserve identity, although image quality is deteriorated (see the bottom of the brain). In addition, CycleGAN creates some artifact inside the pathological region. It is possible that this artifact may indeed be the information that CycleGAN hides to enable input reconstruction. Furthermore, Conditional GAN and vaGAN produce images that are darker and do not match details of the input alluding to possible identity loss. This could be attributed to the lack of losses to help preserve identity, thus making it 'easier' for Conditional GAN and vaGAN to learn a mapping from a pathological to a healthy image of a different subject. Finally, AAE outputs appear blurry and with visible artifacts inside the diseased region.

Quantitative results are shown in Table 4.2, employing now three metrics including one that also assesses deformation correction, as previously described in Section 4.4.4.

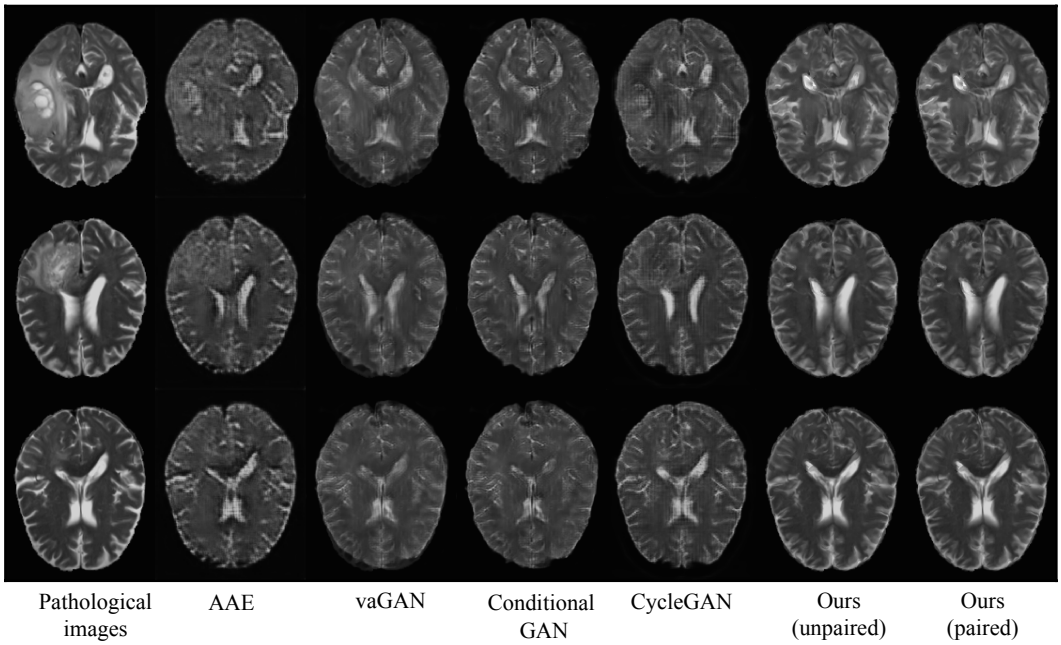


Figure 4.8: Experimental results of three samples, each in every row, for BraTS data. The columns from left to right are the original pathological images, and the synthetic healthy images by *AAE*, *vaGAN*, *Conditional GAN*, *CycleGAN*, and the proposed method in the *unpaired* and *paired* setting, respectively.

As expected, identity of our methods, as measured by iD, has dropped compared to Table 4.1. This is because our methods try to alter the structure of brains to fix the deformations. Indeed, when employing the new metric DeC, our methods achieve higher probability of generated images classified as 'healthy'. For healthiness, h, our methods still outperform the other methods, indicating that the generated images do not contain detectable disease.

4.5.3 Results of expert evaluation on pseudo healthy synthesis for brain tumours

In recognition that our metrics may partially reflect image quality as perceived by expert observers, herein we report the results of our observer study. We aggregated the scores for each approach and averaged across raters to obtain a single consensus score per method per image, for which we used to calculate standard deviation and perform statistical analysis. Given that categorical scores of the human raters and their differences are not normally distributed we instead use a bootstrapped paired t-test [228] to test the null hypothesis described in Section 4.5.1. The human raters consist of members (and previous members) of our group, i.e. Dr. Valerio Giuffrida, Dr. Haochuan Jiang, Dr. Spyridon Thermos, Mr. Grzegorz Jacenkow, Dr. Gabriele Valvano and Mr. Xiao Liu, and an anonymous expert. They all work in medical imaging analysis with years of experience. We also invite Dr. Dafan Yu, a clinical neurologist affiliated with the Third Affiliated Hospital of Sun Yat-sen University, for helping to evaluate the results.

The results of this analysis are shown in Table 4.2. We observe that our methods still outperform baselines and other methods, with a significant improvement for all metrics. In addition, we observe that the methods ranking order is mostly preserved compared to the ranking obtained by the quantitative metrics. Intriguingly, CycleGAN can 'fool' the pre-trained Segmentor which measures healthiness in the 'h' metric but not expert observers in how they assess healthiness. These observations suggest that while numerical evaluation is generally consistent with expert evaluation, there can be room for improvement. We note here the standard deviations for all methods are relatively high, which is due to the binary scoring system used for experiment. Furthermore, we obtained the point biserial correlation [229] between

the values produced by our metrics and the human evaluation study to be 0.35, 0.32, and 0.36 for iD, h, and DeC, respectively. This implies a positive correlation between quantitative and human metrics.

To further evaluate the quality of synthesised images, we requested human observers to discriminate between real and generated 'healthy' images, as described in Section 4.4.4. We calculated the 'realness' score to be 0.43 ± 0.33 for *AAE*, 0.48 ± 0.36 for *vaGAN*, 0.44 ± 0.30 for *Conditional GAN*, 0.47 ± 0.31 for *CycleGAN*, 0.51 ± 0.31 for our method (unpaired), 0.54 ± 0.25 for our method (paired), and 0.63 ± 0.32 for ground-truth healthy images as upper benchmark. Observe that our approaches were the closets to benchmarks.

4.5.4 Segmentation results

Here we evaluate the use of pseudo healthy synthesis on segmentation of T2 BraTS images. Specifically, we compared the pseudo healthy images with the ground-truth pathological images, and obtained the segmentation masks from the difference maps using a threshold of 0.1. For our method, and since segmentation is explicitly performed, we test with masks obtained both from the pseudo healthy images, and from the *Segmentor*. We calculated Dice scores on the test sets to be 0.34 ± 0.11 for AAE, 0.53 ± 0.13 for vaGAN, 0.51 ± 0.14 for conditional GAN, and 0.63 ± 0.16 for CycleGAN. Our approach in the unpaired setting obtained 0.74 ± 0.14 when using the *Segmentor* output, and 0.70 ± 0.13 when using the pseudo healthy images. In both cases our approach achieved statistically significant better results compared to the other benchmarks.

4.5.5 Ablation studies

4.5.5.1 Semi-supervised learning

In this section, we evaluate the effect of the amount of supervision by performing a semi-supervised experiment. Specifically, we vary the number of masks used in the supervised loss of Equation 4.5, while keeping the number of images fixed. The edge cases when all images have paired masks, and vice versa, correspond to the paired and unpaired setting respectively.

Ratio of *paired* samples	0% (unpaired)	20%	40%	60%	80%	100% (paired)
iD	$0.87_{0.04}$	$0.88_{0.05}$	$0.90_{0.06}$	$0.91_{0.05}$	$0.93_{0.04}$	$0.94_{0.03}$
h	$0.88_{0.06}$	$0.87_{0.06}$	$0.89_{0.05}$	$0.88_{0.06}$	$0.89_{0.08}$	$0.89_{0.07}$

Table 4.3: Numerical evaluation of our method on ISLES FLAIR dataset when the ratio of *paired* samples changes. Here x% means that x% of the training pathological images have corresponding ground-truth pathology masks.

Also, the number of segmentation masks used by the unsupervised loss of Equation 4.6 is fixed in all cases. The training strategy is that if the input image has a ground-truth pathology mask, then we use this mask to train the segmentor, with Equation 4.7. When the input image does not have a ground-truth pathology mask, we use the mask adversarial loss to train the network, with Equation 4.8. The results are presented in Table 4.3.

We can observe that for all paired sample ratios, our method can achieve synthetic images of great quality in terms of *identity* and *healthiness*. Nevertheless, we can observe that the *iD*, *i.e.* identity score, increases as the ratio of the paired samples also increases. This could be attributed to the effect of more stable training of the Segmentor. For the ISLES dataset, the Generator needs to learn an identity mapping for healthy regions and a pseudo healthy function for pathological regions. The Segmentor performance has a direct effect on the Reconstructor and an indirect effect on the Generator through back-propagation. With less supervision, the training of the Segmentor is noisier, and the segmented pathological region, that Generator and Reconstructor focus on, is also noisier. Therefore, learning an identity and pseudo healthy function is harder. This affects the identity score, as the Generator must learn to synthesise a whole brain image, and cannot reliably learn an identity function for some parts. On the contrary, the healthiness score, which is directly punished by the adversarial training loss, is not significantly affected. Finally, in order to perform a fair comparison, we trained models at a fixed number of epochs. Even though all models have converged, the noisier training due to the smaller amount of supervision have resulted in a different optimum and therefore the decrease of the identity metric.

4.5.5.2 Unsupervised segmentation and importance of cycle-consistency loss

A pre-requisite for an accurate pseudo healthy synthesis that does not contain traces of pathological information, is for the Segmentor S to be able to accurately extract masks, such that they can be used for the reconstruction of the input pathological images. This should be possible in the unpaired setting as well, where the Segmentor is not trained with any supervision cost. In this setting, the Segmentor is trained using the adversarial loss of the mask discriminator (Equation 4.6), as well as the cycle-consistency loss (Equation 4.3) of the input images.

We evaluate the accuracy of S in the paired and unpaired setting on FLAIR images from ISLES: we obtain respectively an average Dice score of $0.87\ (0.15)$ and $0.79\ (0.17)$ in the testing sets. The results show that even in the unpaired setting, our method can still achieve good segmentation. Results appear to be on par with the numbers provided in [169]. To demonstrate the importance of the cycle-consistency loss (Equation 4.3), we perform an ablation study where we train S only with the adversarial loss of the mask discriminator (*i.e.* only with Equation 4.6). We found that this achieves a Dice of $0.66\ (0.19)$ which is much lower than before. This highlights that just matching the adversary is not enough and that the cycle-consistency loss, by backpropagating additional gradients to the segmentor originating from this cost, encourages further the segmented mask to be correct (in place and size) to enable better reconstruction of the input pathological image.

4.5.5.3 Usefulness and design of Cycle H-H

Our method includes a second training cycle, Cycle H-H, that reconstructs healthy images and masks. This cycle improves the identity preservation of the input images and ensures that our method does not invent disease when a healthy image is given.

Here we perform two ablation studies. For the first ablation study, we train our methods without Cycle H-H, *i.e.* train the network only with Cycle P-H. For the second ablation study, we change Cycle H-H to a new cycle, termed Cycle H-P, which translates healthy images to synthetic diseased ones. The difference between Cycle H-H and Cycle H-P is that Cycle H-H translates a healthy image and a healthy mask to a fake healthy one, and then reconstructs the

Method	iD	h
without Cycle H-H	$0.85_{0.05}$	$0.93_{0.04}$
With Cycle P-H, instead of H-H	$0.89_{0.06}$	$0.89_{0.04}$
Replace Wasserstein with LS-GAN loss	$0.92_{0.03}$	$0.97_{0.04}$
Ours (Cycle H-H & Wasserstein)	$0.94_{0.03}$	$0.99_{0.03}$

Table 4.4: Ablation studies. Here we compare our model with ablated models where we train in the paired setting on ISLES: without Cycle H-H; train with a modified *Cycle H-P* cycle; and also train with Least Square discriminator loss. See text for more details.

input healthy image and mask; while Cycle H-P translates a healthy image and a pathology mask to a fake diseased one, and then reconstructs the input healthy image and pathology mask. The training of Cycle H-P requires an additional discriminator to encourage realistic synthesis of pathological images, and requires careful selection of pathology masks that are suitable to guide the pseudo diseased image generation and fit the real healthy images. We perform the experiments in paired setting on ISLES FLAIR images.

The results are shown in Table 4.4. We observe that our method with Cycle H-H outperforms variants without it and with Cycle H-P. This highlights the importance and effectiveness of the simple, yet effective, design of Cycle H-H in preserving subject identity and improved healthiness of pseudo healthy images.

4.5.5.4 Effectiveness of Wasserstein loss

In this work, to train the discriminators, we replaced the LS-GAN loss [110] that was used in [225], with the Wasserstein loss with gradient penalty [112], which we found to further stabilise training and improve the generated image quality. To illustrate the latter, in Table 4.4 we also show results from models trained in the paired setting on ISLES FLAIR images when using the LS-GAN loss. We observe that Wasserstein loss improves quantitatively the synthetic images in terms of identity and healthiness.

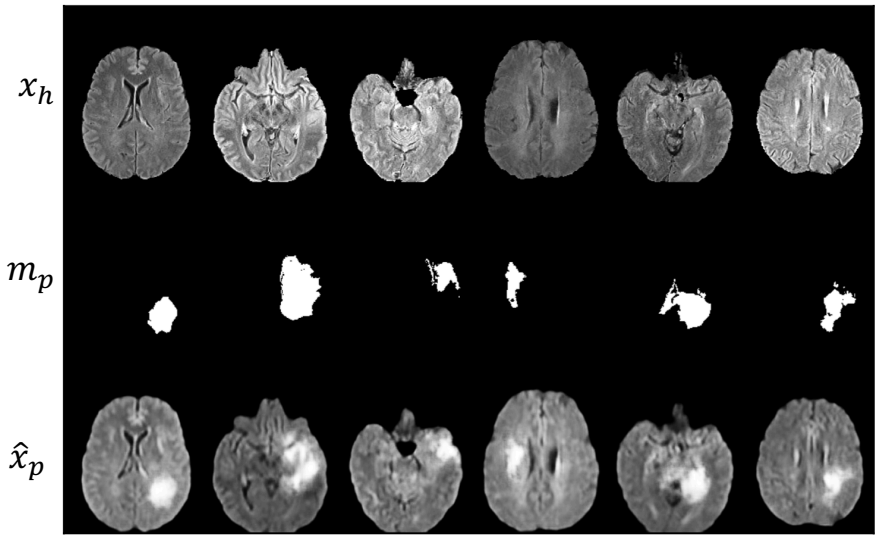


Figure 4.9: Pseudo disease synthesis. Top row shows healthy images, middle row shows random pathology masks, and bottom row presents the synthetic 'pathological' image by the Reconstructor. We can see that Reconstructor can generate realistic 'pathological' images based on input images and masks.

4.5.5.5 Pseudo disease synthesis

If our method works well, the Reconstructor should be able to synthesise a 'pathological' image given a healthy one and a suitable pathology mask. Here we used randomly sampled but suitable masks and healthy images to generate a pathological image, with a trained Reconstructor on the ISLES FLAIR dataset. We show some example images of this pseudo disease synthesis, as shown in Figure 4.9. We can observe that although our model has never been trained to perform this pseudo disease synthesis, the Reconstructor can synthesise a 'pathological' image when given a healthy image and a suitable pathology mask.

4.6 Summary

This chapter proposed a method that aims to synthesise pseudo healthy images using an adversarial design that disentangles pathology. Our method is composed of a Generator that creates pseudo healthy images and a Segmentor that predicts a pathology map. These key components are trained aided by the Reconstructor, which reconstructs the input pathological

image conditioned on the map and the pseudo healthy image. Our method can be trained using supervised and adversarial loses taking advantage of unpaired data. We propose numerical evaluation metrics to explicitly measure the quality of the synthesised images. We demonstrate on ISLES, BraTS and Cam-CAN datasets that our method outperforms baselines both qualitatively, quantitatively, and subjectively with a human study. In this chapter, we measure the reliability of generated images with quantitative metrics and human evaluation. However, we admit that the golden standard of measuring the reliability should be its utility in practical applications, which is seen as a future work. We do not explicitly consider the variability of the generated images as we assume that the 'healthy' version of a pathological image should be unique. However, as a future direction, we could compare the variability of all pathological images and their pseudo healthy counterfactuals. We expect that there would a reduction of variability in the pseudo healthy images, due to the removal of pathology. From a conditioning perspective, the proposed approach of this chapter treated 'diseased' or 'healthy' as discrete factors and generated images based on these discrete factors. However, in some cases we might need to generate images conditioned on continuous factors, *e.g.* age, which is a more difficult task. In Chapter 5, we will present a new approach for brain ageing synthesis.

Chapter 5
Brain Ageing Synthesis[1]

5.1 Introduction

In this chapter, we focus on a complex image synthesis task, *i.e.* synthesising the visual predictions of brain images with age as a factor. Similar to Chapter 4, we adopt adversarial training to generate 'older' images given baseline observations. Since longitudinal brain data is limited, we use cross-sectional data to train the model. The proposed model is able to predict subject-specific future images conditioned on age and health state.

The ability to predict the future state of an individual can be of great benefit for longitudinal studies [198]. However, such learned phenomenological predictive models need to capture anatomical and physiological changes due to ageing and separate the factors that influence future state. Recently, deep generative models have been used to simulate and predict future degeneration of a human brain using existing scans [191, 188, 189]. However, current methods require considerable amount of longitudinal data to sufficiently approximate an auto-regressive model. Here, we propose a new conditional adversarial training procedure that does *not* require longitudinal data to train. Our approach (shown in Fig. 5.1) synthesises images of aged brains for a desired age and health state.

Brain ageing, accompanied by a series of functional and physiological changes, has been intensively investigated [230, 231]. However, the underlying mechanism has not been com-

[1] **This chapter is based on the following publications:**

- **Tian Xia, Agisilaos Chartsias, and Sotirios A. Tsaftaris, for the Alzheimer's Disease Neuroimaging Initiative. "Consistent Brain Ageing Synthesis". In: Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11767. Springer, Cham. https://doi.org/10.1007/978-3-030-32251-9-82.**
- **Tian Xia, Agisilaos Chartsias, Chengjia Wang, Sotirios A. Tsaftaris. "Learning to synthesise the ageing brain without longitudinal data", Medical Image Analysis, Volume 73, 2021, 102169, ISSN 1361-8415.**

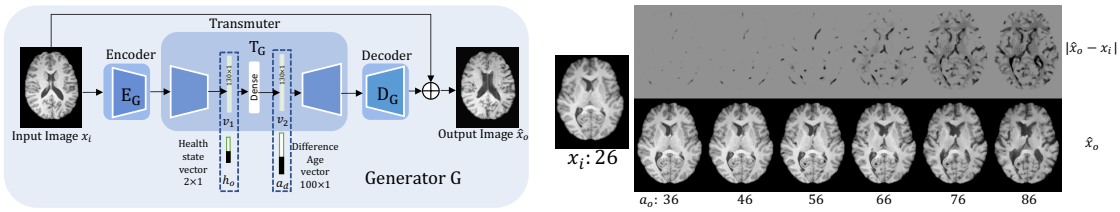


Figure 5.1: A schematic of proposed method and example results for an image. **Left:** The input is a brain image x_i, and the network synthesises an aged brain image $\hat{x}_o$ from x_i, conditioned on the target health state vector h_o and target age difference $a_d = a_o - a_i$ between input a_i and target a_o ages, respectively. **Right:** For an image x_i of a 26 year old subject, bottom row shows outputs $\hat{x}_o$ given different target age. The top row shows the corresponding image differences $|\hat{x}_o - x_i|$ to highlight progressive changes.

pletely revealed [39, 232]. Prior studies have shown that a brain's chronic changes are related to different factors, *e.g.* the biological age [47], degenerative diseases such as Alzheimer's Disease (AD) [233], binge drinking [234], and even education [235]. Accurate simulation of this process has great value for both neuroscience research and clinical applications to identify age-related pathologies [232, 47].

One particular challenge is inter-subject variation: every individual has a unique ageing trajectory. Previous approaches built a spatio-temporal atlas to predict average brain images at different ages [178, 177]. However, individuals with different health status follow different ageing trajectories. An atlas may not preserve subject-specific characteristics; thus, may preclude accurate modelling of individual trajectories and further investigation on the effect of different factors, *e.g.* age, gender, education, etc [192]. Recent studies proposed subject-specific ageing progression with neural networks [191, 188], although they require longitudinal data to train. Ideally, the longitudinal data should cover a long time span with frequent sampling to ensure stable training. However, such data are difficult and expensive to acquire, particularly for longer time spans. Even in ADNI [5], one of the most well-known large-scale datasets, longitudinal images are acquired at few time points and cover only a few years. Longitudinal data of sufficient time span remain an open challenge.

In this chapter, we build the foundations of a model that can be trained without longitudinal data. A simplified schematic of our model is shown in Fig. 5.1 along with example results.

Given a brain image, our model produces a brain of the same subject at target age. The input image is first encoded into a latent space, which is modulated by two vectors representing target age difference and health state (AD status in this chapter), respectively. The conditioned latent space is finally decoded to an output image, *i.e.* the synthetically aged image.

Under the hood, what trains the generator, is a deep adversarial method that learns the joint distribution of brain appearance, age and health state. The quality of the output is encouraged by a discriminator that judges whether an output image is representative of the distribution of brain images of the desired age and health state. A typical problem in synthesis which is exacerbated with *cross-sectional* data [198] is loss of *subject identity*[2], *i.e.* the synthesis of an output that may not correspond to the input subject's identity. We propose, and motivate, two loss functions towards retaining *subject identity* by regularising the amount of change introduced by ageing. In addition, we motivate the design of our conditioning mechanisms and show that ordinal binary encoding for both discrete and continuous variables improves performance significantly.

We consider several metrics and evaluation approaches to verify the quality and biological plausibility of our results. We quantitatively evaluate our simulation results using longitudinal data from the ADNI dataset [5] with classical metrics that estimate image fidelity. Since the longitudinal data only cover a limited time span, it is difficult to evaluate the quality of synthesized aged images across decades. For brain ageing synthesis, a good synthetic brain image should be accurate in terms of age, *i.e.* be close to the target age that we want it to be, and also preserve subject identity, *i.e.* should be from the same subject as the input. Thus, we pre-train a deep network to estimate the apparent age from output images. The estimated ages are used as a proxy metric for *age accuracy*. We also show qualitative results, including ageing simulation on different health states and long-term ageing synthesis. Both

[2]A classical computer vision example is generating a human face resembling another individual instead of the input subject. Even with faces, humans find it difficult to assess identity loss. It remains hard to define detailed structural changes during ageing, *e.g.* balding, nose shape change, eye colour change. There are some common patterns that we can expect, such as wrinkles and gray/white hair, but it is difficult to define other more detailed changes. Therefore, even in face ageing, 'subject identity' is defined as young and old images should be from the same person. In brain synthesis, it is even more difficult to define 'subject identity', as human eyes are less able to visually ascertain brain image identity particularly as modulated by age and pathology. In this chapter, we followed a similar analogue of 'identity': a "synthetic image should be from the same subject as the input image".

quantitative and qualitative results show that our method outperforms benchmarks with more accurate simulations that capture the characteristics specific to each individual on different health states. Furthermore, we train our model on Cam-CAN and evaluate it on ADNI to demonstrate the generalisation ability to unseen data. In addition, to demonstrate the realism of synthetic results, we perform volume synthesis and evaluate deformation. We also estimate gray matter atrophy in middle temporal gyrus and find that our model, even without longitudinal data, has learned that ageing and disease leads to atrophy. Ablation studies investigate the effect of loss components and different ways of embedding clinical variables into the networks.

Our contributions are summarised as follows:

- Our main contribution is a deep learning model that learns to simulate the brain ageing process, and perform subject-specific brain ageing synthesis, trained on *cross-sectional* data overcoming the need for longitudinal data.
- For our model to be able to change output based on desired input (age and health state), we use an (ordinal) embedding mechanism that guides the network to learn the joint distribution of brain images, age and health state.
- Since we do not use longitudinal data that can constrain the learning process, we design losses that aim to preserve subject identity, while encouraging quality output.
- We provide an experimental framework to verify the quality and biological validity of the synthetic outputs.

While our first contribution is the most important one, it is the combination of our proposed losses and embedding mechanisms that lead to the method's robustness, as extensive experiments and ablation studies on two publicly available datasets, namely Cam-CAN [2] and ADNI [5] show.

The manuscript proceeds as follows: Section 5.2 reviews related work on brain ageing simulation and prediction. Section 5.3 details the proposed method. Section 5.4 describes the experimental setup and training details. Section 5.5 presents results and discussion. Finally, Section 5.6 offers summary.

5.2 Related Work

Here we first summarise related works for *brain ageing simulation*, *i.e.* simulating the ageing process from data. *Please refer to Section 3.3.2 for a more detailed review*. For completeness, we also briefly discuss *brain age prediction*, *i.e.* estimating age from an image.

5.2.1 Brain ageing simulation

Given variables such as age, one can synthesise the corresponding brain image to enable visual observation of brain changes. For instance, patch-based dictionary learning [173], kernel regression [177, 198, 179], linear mixed-effect modelling [180, 181] and non-rigid registration [182, 183, 184, 185] have been used to build spatio-temporal atlases of brains at different ages. However, by relying on population averages as atlases, subject-specific ageing trajectories are harder to capture.

Deep generative methods have also been used for this task. While [188, 189] and [187] used formulations of Generative Adversarial Networks (GAN) [107] to simulate brain changes, others [191] used a conditional adversarial autoencoder as the generative model, following a recent face ageing approach [193]. Irrespective of the model, these methods need longitudinal data, which limits their applicability. To allow training with cross-sectional data, GAN-based or VAE-based models have been used in [194, 195, 196, 197]. However, they either modelled the brain ageing a linear process [194], or resulted in blurry results without quantitative evaluation [195, 196, 197].

In summary, most previous methods either built average atlases [173, 177, 198, 179], or required longitudinal data [188, 189, 191, 187] to simulate brain ageing. Other methods either did not consider subject identity [194, 195], or did not evaluate in detail morphological changes [196, 197]. To address these shortcomings, we propose a conditional adversarial training procedure that learns to simulate the brain ageing process by being *specific* to the input subject, and by learning from *cross-sectional* data *i.e.* without requiring longitudinal observations.

5.2.2 Brain age prediction

These methods predict age from brain images learning a relationship between image and age; thus, for completeness we briefly mention two key directions. For example, [236] predicted age with hand-crafted features and kernel regression whereas [237] used Gaussian Processes. Naturally performance relies on the effectiveness of the hand-crafted features.

Recently, deep learning models have been used to estimate the brain age from imaging data. For example, [238] used a VGG-based model [239] to predict age and detect degenerative diseases, while [240] proposed to discover genetic associations with the brain degeneration using a ResNet-based network [13]. Similarly, [241] used a CNN-based model to predict age. [242] used the age predicted by a deep network to detect traumatic brain injury. While most previous works achieved mean absolute error (MAE) of 4-5 years, [241] achieved state-of-the-art performance with MAE of 2.14 years. However, these methods did not consider the morphological changes of brain, which is potentially more informative [243].

5.3 Proposed approach

5.3.1 Problem statement, notation and overview

In the rest of the chapter, we use **bold** notations for vectors/images, and *italics* notations for scalars. For instance, a represents an age while $\mathbf{a}$ is a vector that represents age a. We denote a brain image as $\mathbf{x}_s$ (and $\mathcal{X}_s$ their distribution such that $\mathbf{x}_s \sim \mathcal{X}_s$), where s are the subject's clinical variables including the corresponding age a and health state (AD status) h. Given a brain image $\mathbf{x}_i$ of age a_i and health state h_i, we want to synthesise a brain image $\hat{\mathbf{x}}_o$ of target age a_o and health state h_o. Critically, the synthetic brain image $\hat{\mathbf{x}}_o$ should retain the subject identity, *i.e.* belong to the same subject as the input $\mathbf{x}_i$, throughout the ageing process. The contributions of our approach, shown in Fig. 5.2, are the design of the conditioning mechanism; our model architecture that uses a Generator to synthesise images, and a Discriminator to help learn the joint distribution of clinical variables and brain appearance; and the losses we use to guide the training process. We detail all these below.

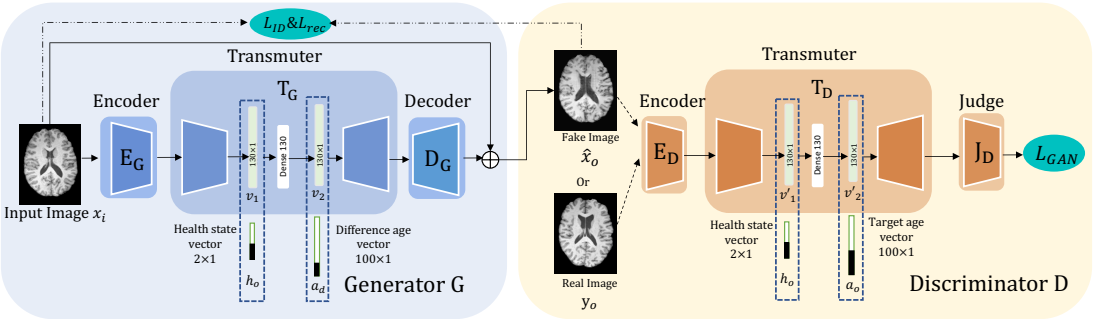


Figure 5.2: An overview of the proposed method (training). $\mathbf{x}_i$ is the input image; $\mathbf{h}_o$ is the target health state; $\mathbf{a}_d$ is the difference between the starting age a_i and target age a_o: $a_d = a_o - a_i$; $\hat{\mathbf{x}}_o$ is the output (aged) image (supposedly belong to the same subject as x_i) of the target age a_o and health state h_o. The *Generator* takes as input $\mathbf{x}_i$, $\mathbf{h}_o$ and $\mathbf{a}_d$, and outputs $\hat{\mathbf{x}}_o$; the *Discriminator* takes as input a brain image and $\mathbf{h}_o$ and $\mathbf{a}_o$, and outputs a discrimination score.

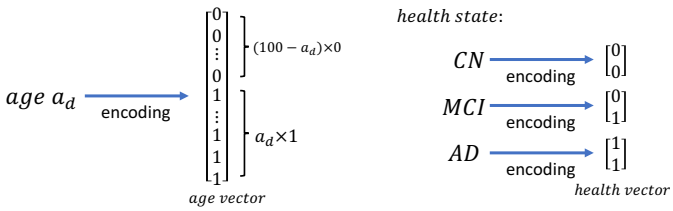


Figure 5.3: Ordinal encoding of age and health state. **Left** shows how we represent age $\mathbf{a}_d$ using a binary vector with first $\mathbf{a}_d$ elements as 1 and the rest as 0; **Right** is the encoding of health state, where we use a 2×1 vector to represent three categories of AD status: control normal (CN), mildly cognitive impaired (MCI), and Alzheimer's Disease (AD).

5.3.2 Conditioning on age and health state

As most previous works, we simulate the ageing brain with age as a factor. However, brain ageing is not only affected by age, but also by other clinical factors such as neurodegenerative disease. Here, we also involve the health state, *i.e.* AD status, as another factor to better simulate the ageing process. [3]

[3] Additional fine-grained information on AD effects on different, local, brain regions could be provided if clinical scores are used instead. As our work is the first to attempt to learn without longitudinal data, for simplicity we focused on variables capturing global effects. In the summary section, we note the addition of fine-grained information as an avenue for future improvement.

We use ordinal binary vectors, instead of one-hot vectors as in [193], to encode both age and health state, which are embedded in the bottleneck layer of the Generator and Discriminator (detailed in Section 5.3.4). We assume a maximal age of 100 years and use a 100×1 vector to encode age a. Similarly, we use a 2×1 vector to encode health state. A simple illustration of this encoding is shown in Fig. 5.3. An ablation study presented in Section 5.5.4 illustrates the benefits of *ordinal* v.s. *one-hot* encoding.

5.3.3 Preliminary method: brain ageing only conditioned on age

Here we first introduce our preliminary method [15], where we only condition brain ageing synthesis on age without considering health state.

Model: The preliminary model consists of a *Generator* and a *Discriminator*, shown in Fig. 5.4. These are detailed below. Note images x_i, $\hat{x}_o$ and y_o represent the input, the aged output and a real older brain image from another subject, respectively.

Generator: 'G' takes as input a 2D brain image x_o and an ordinal age vector representing the age difference between a_o and a_i:$a_d = a_o - a_i$, and outputs a 2D older image $\hat{x_o}$. We condition on a_d such that when input and output ages are equal ($a_d = 0$) the network is drawn to recreate its input. This works in synergy with our identity-preserving loss described below.

The *Generator* consists of three subnetworks: '*Encoder*' E_G , '*Transmuter* 'T_G[4] , and '*Decoder*' D_G. E_G extracts latent features F_{e1} from input x_{t_i}: $F_{e1} = E_G(x_i)$. T_G outputs a feature map $F_{e2} = T_G(F_{e1}, v_d)$ by first transforming F_{e1} to a bottleneck vector v_1, and by concatenating c_{e1} with a_d. To keep networks parameters low we empirically set the size of v_1 to *130*. Afterwards, to preserve information of x_i, and achieve accurate synthetic results, we introduce a skip connection between F_{e1} and F_{e2}: $F_{e3} = cat(F_{e1}, F_{e2})$, where cat($\cdot$) concatenates the elements of the given tensors along the channel dimension. Finally, the Decoder D_G synthesises the aged output $\hat{x}_t$ from F_{e3}. $\hat{x}_o$ should manifest the characteristics of brains at age t_o whilst preserving the identity of input x_i, *i.e.* $\hat{x}_o$ should be the brain image of the same subject as x_i at age t_o.

[4]We change the name of this subnetwork to 'Transmuter' to avoid confusion between it and the popular model Transformer used in NLP.

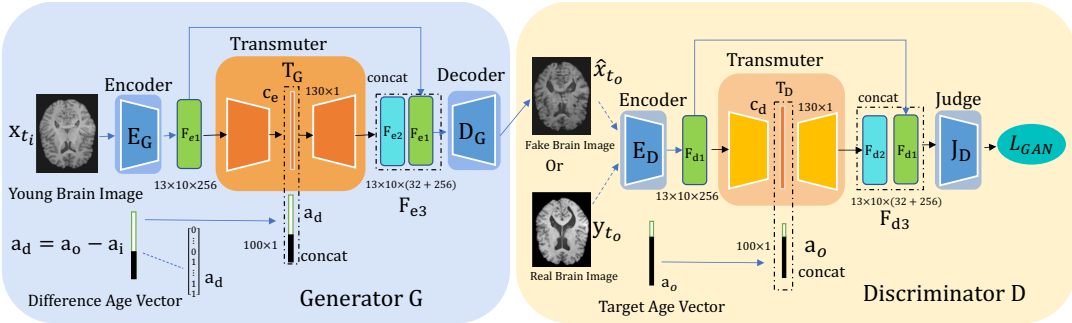


Figure 5.4: Preliminary method. x_i is the input image of age a_i; $\hat{x}_o$ is the output (aged) image (supposedly of the same subject as x_i) at the age a_o; a_o is the target age vector and a_d is the difference age vector corresponding to $a_d = a_o - a_i$. The *Generator* takes as input x_i and a_d, and outputs $\hat{x}_o$; the *Discriminator* takes as input an image and a target age vector, and outputs a Wasserstein score.

Discriminator: D contains an Encoder E_D and a Transmuter T_D to condition on target age and a Judge J_D to output a discriminator score. Note here we condition on a_o, instead of a_d, to learn the joint distribution of brain appearance and age, such that it can discriminate real vs. synthetic images of correct age.

To summarise, the forward pass for the *Generator* is $\hat{x}_t = G(x_i, a_d)$, and for the *Discriminator* is $w_{fake} = D(\hat{x}_o, a_o)$ and $w_{real} = D(y_o, a_o)$.

Losses: The overall training loss is defined as:

$$\mathcal{L} = \max_G \min_D \mathcal{L}_{GAN} + \min_G \lambda_{ID} \mathcal{L}_{ID},$$

where $\mathcal{L}_{GAN}$ is the GAN loss, and $\mathcal{L}_{ID}$ is an *age-modulated identity-preserving loss* and $\lambda_{ID} = 100$ the weight of $\mathcal{L}_{ID}$. $\mathcal{L}_{GAN}$ pushes the solution towards realistic images of correct age, whereas $\mathcal{L}_{ID}$ pushes towards subject-specific synthesis.

$\mathcal{L}_{GAN}$ is a Wasserstein loss with gradient penalty for stable training [112]:

$$\mathcal{L}_{GAN} = \mathbb{E}_{y_o \sim \mathcal{X}_o, \hat{x}_o \sim \hat{\mathcal{X}}_o}[D(y_o, a_o - D(\hat{x}_o, a_o) + \lambda_{GP}(\|\nabla_{\tilde{z}} D(\tilde{z})\|_2 - 1)^2],$$

where $\tilde{z}$ is the average sample defined as $\tilde{z} = \epsilon \hat{x}_o + (1 - \epsilon) y_o$, $\epsilon \sim U[0, 1]$. First two terms

measure the Wasserstein distance between real and fake samples; last term is the gradient penalty. As in [112] we set $\lambda_{GP} = 10$.

Although the preliminary method achieve realistic brain synthesis according to given ages, it does not consider other factors that can affect brain ageing progression. Therefore, in Section 5.3.4, we extend it by involving AD state as another factor as our proposed method for this chapter. The preliminary method is compared to the proposed method as a benchmark method in Section 5.5. We denote this preliminary method as *ours-previous*.

5.3.4 Proposed model

The preliminary method introduced in Section 5.3.3 only considers age, but other factors such as health status also affect the appearance of brains. Here we extend the preliminary method by involving AD state as another factor.

The proposed method consists of a *Generator* and a *Discriminator*. The Generator synthesises aged brain images corresponding to a target age and a health state. The Discriminator has a dual role: firstly, it discriminates between ground-truth and synthetic brain images; secondly, it ensures that the synthetic brain images correspond to the target clinical variables. The Generator is adversarially trained to generate realistic brain images of the correct target age. The detailed network architectures are shown in Fig. 5.5.

5.3.4.1 Generator

Here, the Generator is similar to that of the preliminary method in Section 5.3.3. The difference lies in the latent space, where in addition to age vector a_d, we also involve the health vector h_o as another conditioning factor. Details are described below.

The Generator G takes as input a 2D brain image $\mathbf{x}_i$, and ordinal binary vectors for target health state h_o and age difference a_d. Here, we condition on the age difference between input age a_i and target age a_o: $a_d = a_o - a_i$, such that when input and output ages are equal $a_d = 0$, the network is encouraged to recreate the input. The output of G is a 2D brain image $\hat{\mathbf{x}}_o$

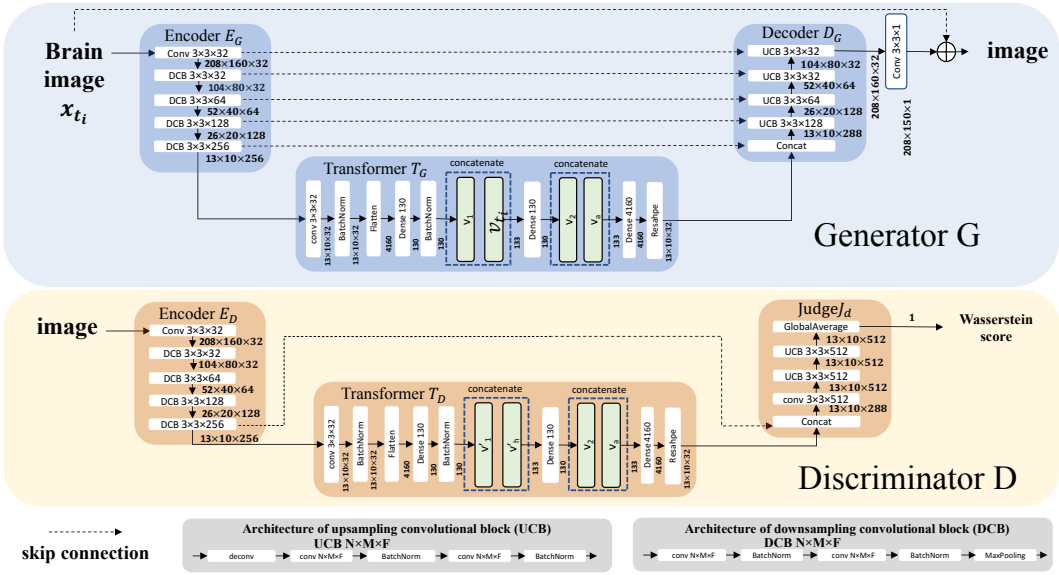


Figure 5.5: Detailed architectures of *Generator* and *Discriminator*. The Generator contains three parts: an Encoder to extract latent features; a Transmuter to involve target age and health state; and a Decoder to generate aged images. Similarly, we use the same conditioning mechanism for the Discriminator to inject the information of age and health state, and a long skip connection to better preserve features of input image.

corresponding to the target age and health state.[5]

G has three components: the *Encoder* E_G, the *Transmuter* T_G and the *Decoder* D_G. E_G first extracts latent features from the input image $\mathbf{x}_i$; T_G involves the target age and health state into the network. Finally, D_G generates the aged brain image from the bottleneck features. To embed age and health state into our model, we first concatenate the latent vector $\mathbf{v}_1$, obtained by E_G, with the health state vector h_o. The concatenated vector is then processed by a dense layer to output latent vector $\mathbf{v}_2$, which is then concatenated with the difference age vector $\mathbf{a}_d$. Finally, the resulting vector is used to generate the output image.[6] We adopt long-skip connections [14] between layers of E_G and D_G to preserve details of the input image

[5]Note that the target health state can be different from the corresponding input state. This encourages learning a joint distribution between brain images and clinical variables.

[6]We tested the ordering of $\mathbf{h}_o$ and $\mathbf{a}_d$, and it did not affect the results. We also tried to concatenate $\mathbf{h}_o$, $\mathbf{a}_d$ and $\mathbf{v}_1$ together into one vector, and use the resulting vector to generate the output. However, we found that the model tended to ignore the information of $\mathbf{h}_o$. This might be caused by the dimensional imbalance between h_o (2×1) and a_d (100×1).

and improve the sharpness of the output images. Overall, the Generator's forward pass is: $\hat{\mathbf{x}}_o = G(\mathbf{x}_i, \mathbf{a}_d, \mathbf{h}_o)$.

5.3.4.2 Discriminator

Similar to the Generator, the Discriminator D contains three subnetworks: the *Encoder* E_D that extracts latent features, the *Transmuter* T_D that involves the conditional variables, and the *Judge* J_D that outputs a discrimination score. For the discriminator to learn the joint distribution of brain image, age, and health state, we embed the age and health vectors into the discriminator with a similar mechanism as that of the Generator.

Note that D is conditioned on the target age $\mathbf{a}_o$ instead of age difference $\mathbf{a}_d$, to learn the joint distribution of brain appearance and age, such that it can discriminate between real and synthetic images of correct age. The forward pass for the Discriminator is $w_{fake} = D(\hat{\mathbf{x}}_o, \mathbf{a}_o, \mathbf{h}_o)$ and $w_{real} = D(\mathbf{y}_o, \mathbf{a}_o, \mathbf{h}_o)$.

5.3.5 Losses

We train with a multi-component loss function containing *adversarial*, *identity-preservation* and *self-reconstruction* losses. We detail these below.

5.3.5.1 Adversarial loss

We adopt the Wasserstein loss with gradient penalty [112] to predict a realistic aged brain image $\hat{\mathbf{x}}_o$ and force $\hat{\mathbf{x}}_o$ to correspond to the target age a_o and health state h_o:

$$\begin{aligned} L_{GAN} = \mathbb{E}_{\mathbf{y}_o \sim \mathcal{X}_o, \hat{\mathbf{x}}_o \sim \hat{\mathcal{X}}_o}[D(\mathbf{y}_o, \mathbf{a}_o, \mathbf{h}_o) \\ D(\hat{\mathbf{x}}_o, \mathbf{a}_o, \mathbf{h}_o) + \lambda_{GP}(\|\nabla_{\tilde{z}} D(\tilde{\mathbf{z}}, \mathbf{a}_o, \mathbf{h}_o)\|_2 - 1)_2], \end{aligned} \tag{5.1}$$

where $\hat{\mathbf{x}}_o$ is the output image: $\hat{\mathbf{x}}_o = G(\mathbf{x}_i, \mathbf{a}_d, \mathbf{h}_o)$ (and $\mathbf{a}_d = \mathbf{a}_o - \mathbf{a}_i$); $\mathbf{y}_o$ is a ground truth image from another subject of target age a_o and health state h_o; and $\tilde{\mathbf{z}}$ is the average sample defined by $\tilde{\mathbf{z}} = \epsilon \hat{\mathbf{x}}_o + (1 - \epsilon)\mathbf{y}_o, \epsilon \sim U[0, 1]$. The first two terms measure the Wasserstein

distance between ground-truth and synthetic samples; the last term is the gradient penalty involved to stabilise training. As in [112] and [12] we set $\lambda_{GP} = 10$.

5.3.5.2 Identity-preservation loss

While L_{GAN} encourages the network to synthesise realistic brain images, these images may lose subject identity. For example, it is easy for the network to learn a mapping to an image that corresponds to the target age and health state, but belongs to a different subject. An illustration is presented in Fig. 5.6, where ageing trajectories of two subjects are shown. The task is to predict the brain image of subject 1 at age a_2 starting at age a_1, by learning a mapping from point A to point B. But there are no ground-truth data to ensure that we stay on the trajectory of subject 1. Instead, the training data contain brain images of age a_2 belonging to subject 2 (and other subjects). Using only L_{GAN}, the Generator may learn a mapping from A to C to fool the Discriminator, which will lose the identity of subject 1. To alleviate this and encourage the network to learn mappings along the trajectory (*i.e.* from A to B), we adopt:

$$L_{ID} = \mathbb{E}_{\mathbf{x}_i \sim \mathcal{X}_i, \hat{\mathbf{x}}_o \sim \mathcal{X}_o} \left\| \mathbf{x}_i - \hat{\mathbf{x}}_o \right\|_1 \cdot e^{-\frac{|a_o - a_i|}{|a_{max} - a_{min}|}}, \tag{5.2}$$

where $\mathbf{x}_i$ is the input image of age a_i and $\hat{\mathbf{x}}_o$ is the output image of age a_o ($a_o > a_i$). The term $e^{-\frac{|a_o - a_i|}{|a_{max} - a_{min}|}}$ encourages $\left\| \mathbf{x}_i - \hat{\mathbf{x}}_o \right\|_1$ to positively correlate with the difference $|a_o - a_i|$. The health state is not involved in L_{ID} as we do not aim to precisely model the ageing trajectory.

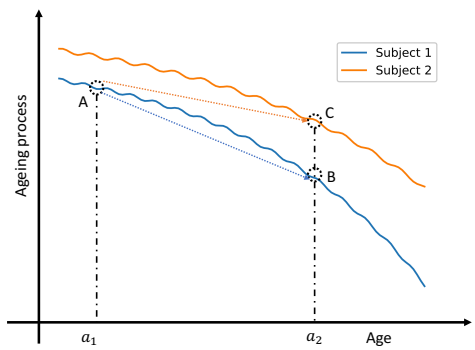


Figure 5.6: Illustration of ageing trajectories for two subjects. For a subject of age a_1 (A), the network can learn a mapping from A to C, which could still fool the Discriminator, but loses the identity of Subject 1 (orange line).

Instead, L_{ID} is used to encourage identity preservation by penalising major changes between images close in age, and to stabilise training. A more accurate ageing prediction, which is also correlated with health state, is achieved by the adversarial loss. An ablation study illustrating the critical role of L_{ID} is included in Section 5.5.4.

5.3.5.3 Self-reconstruction loss

We use a self-reconstruction loss,

$$L_{rec} = \mathbb{E}_{\mathbf{x}_i \sim \mathcal{X}_i, \hat{\mathbf{x}}_o \sim \mathcal{X}_i} \left\| \mathbf{x}_i - \hat{\mathbf{x}}_o \right\|_1, \tag{5.3}$$

to explicitly encourage that the output $\hat{\mathbf{x}}_o$ is a faithful reconstruction of the input $\mathbf{x}_i$ for the same age and health state. Although L_{rec} is similar to L_{ID}, their roles are different: L_{ID} helps to preserve subject identity when generating aged images, while L_{rec} encourages smooth progression via self-reconstruction. An ablation study on L_{rec} in Section 5.5.4 shows the importance of stronger regularisation.

5.4 Experimental setup

Datasets: For the experiments of this chapter, we use Cam-CAN and ADNI datasets, which are introduced in Section 2.5. Here we briefly introduce the modalities and numbers of volumes used for training and evaluation.

Cambridge Centre for Ageing and Neuroscience (Cam-CAN) [2] is a cross-sectional dataset containing normal subjects aged 18 to 88. We split subjects into different age groups spanning 5 years. We randomly selected 38 volumes from each age group and used 30 for training and 8 for testing. To prevent data imbalance, we discarded subjects under 25 or over 85 years old, because there are underrepresented data. We use Cam-CAN to demonstrate consistent brain age synthesis across the whole lifespan. *Alzheimer's Disease Neuroimaging Initiative (ADNI)* [5] is a longitudinal dataset. We use ADNI to demonstrate brain image synthesis conditioned on different health states. Since ADNI has longitudinal data, we used these data

to quantitatively evaluate the quality of synthetically aged images. We chose 786 subjects as training (279 CN, 260 MCI, 247 AD), and 136 subjects as testing data (49 CN, 46 MCI, 41 AD). The age difference between baseline and followup images in the testing set is 2.93 ± 1.35 years.

Pre-processing: All volumetric data are skull-stripped using DeepBrain[7], and linearly registered to MNI 152 space using FSL-FLIRT [94]. We normalise brain volumes by clipping the intensities to $[0, V_{99.5}]$, where $V_{99.5}$ is the 99.5% largest intensity value within each volume, and then rescale the resulting intensities to the range $[-1, +1]$. Such intensity pre-processing also helps alleviate potential intensity harmonisation issues between datasets in a manner that creates no leakage (see footnote on section 5.2.3 why this is important). We select the middle 60 axial slices from each volume, and crop each slice to the size of $[208, 160]$. Note the training and testing are performed on these selected 2D slices, instead of all available 2D slices, unless specified. During training, we only use *cross-sectional* data, *i.e.* one subject only has one volume of a certain age. During testing, we use the longitudinal ADNI data covering more than 2 years, and discard data where images are severely misaligned due to registration errors.

Benchmarks: We compare with the following benchmarks[8]:

Conditional GAN: We use a conditional image-to-image translation approach [122] and train different Conditional GANs for transforming young images to different older age groups. Therefore, a single model of ours is compared with age-group specific Conditional GANs.
CycleGAN: We use CycleGAN [130], with two translation paths: from 'young' to 'old' to 'young', and from 'old' to 'young' to 'old'. Similarly to Conditional GAN, we train several CycleGANs for different target age groups.

CAAE: We compare with [193], a recent paper for face ageing synthesis. We use the official implementation[9], modified to fit our input image shape. This method used a Conditional

[7]https://github.com/iitzco/deepbrain

[8]We also used the official implementation of [195]; however, our experiments confirmed the poor image quality reported by the author.

[9]https://zzutk.github.io/Face-ageing-CAAE/

Adversarial Autoencoder (CAAE) to perform face ageing synthesis by concatenating a one-hot age vector with the bottleneck vector. They divided age into discrete age groups.

Ours-previous: We also compare with our preliminary method [15], described in Section 5.3.3.

Implementation details: The optimization function is:

$$L = \min_{\mathrm{G}} \max_{\mathrm{D}} \lambda_1 L_{GAN} + \lambda_2 L_{ID} + \lambda_3 L_{rec}, \tag{5.4}$$

where $\lambda_1 = 1$, $\lambda_2 = 100$ and $\lambda_3 = 10$ are hyper-parameters used to balance each loss. The λ parameters are chosen experimentally. We chose λ_2 as 100 following [12] and [15], and λ_3 as a smaller value to put emphasis on synthesis rather than self-reconstruction.

To train our model, we divide subjects into a young group and an old group, and randomly draw an image $\mathbf{x}_i$ the young group and an image $\mathbf{y}_o$ from the old group to synthesise the aged image $\hat{\mathbf{x}}_o$ (of $\mathbf{x}_i$) with target age a_o and health state h_o (of those corresponding to $\mathbf{y}_o$). Here $\hat{\mathbf{x}}_o$ is the synthetically aged version of $\mathbf{x}_i$, and the target age a_o and health state h_o are the same as those of the selected old sample $\mathbf{y}_o$. Afterwards, $\mathbf{y}_o$ and $\hat{\mathbf{x}}_o$ are fed into the discriminator as real and fake samples, respectively. Note that for all samples $a_o > a_i$, and h_o could be different than h_i. Since Alzheimer's Disease is an irreversible neurodegenerative disease, we select samples where the input health status is not worse than the output health status. We train all methods for 600 epochs. We update the generator and discriminator iteratively [111, 107]. Since the discriminator of a Wasserstein GAN needs to be close to optimal during training, we update the discriminator for 5 iterations per generator update. Initially, for the first 20 epochs, we update the discriminator for 50 iterations per generator update. We use Keras [223] and train with Adam [224] with a learning rate of 0.0001 and decay of 0.0001. Code is available at `https://github.com/xiat0616/BrainAgeing`.

Evaluation metrics: To evaluate the quality of synthetically aged images, we use the longitudinal data from ADNI dataset. We select follow-up studies covering >2 years to allow observable neurodegenerative changes to happen. We used standard definitions of *mean squared error* (MSE), *peak signal-to-noise ratio* (PSNR) and *structural similarity* (SSIM) of window length of 11 [206] to evaluate the closeness of the predicted images to the ground-truth.

Predicted age difference (PAD) as a metric: Longitudinal data in ADNI only cover a short time span, *i.e.* the age difference between baseline and followup images is only several years. To assess output even when we do not have corresponding follow-up ground truth, we use a proxy metric of apparent age to evaluate image output. To develop our proxy metric, we first train a learning based age predictor to assess apparent brain age. We pre-train a VGG-like [239] network to predict age from brain images, then use this age predictor, f_{pred}, to estimate the apparent age of output images. To train this age predictor f_{pred} we combine Cam-CAN and healthy (CN) ADNI training data to ensure good age coverage. On a held out testing set it achieves a MAE of 5.1 ± 3.1 years. When the held out dataset is restricted to ADNI healthy subjects alone, MAE is 3.9 ± 2.8 years.

We use the difference between the predicted and desired target age to assess how close the generated images are to the (desired) target age. Formally, our proxy metric *predicted age difference* (PAD) is:

$$PAD = \mathbb{E}_{\hat{\mathbf{x}}_o \sim \mathcal{X}_o} |f_{pred}(\hat{\mathbf{x}}_o) - a_o|, \tag{5.5}$$

where f_{pred} is the trained age predictor, $\hat{\mathbf{x}}_o$ is the synthetically aged image, and a_o is the target age. Here we choose to measure the mean absolute error as we want to avoid the neutralization of positive and negative errors. By adopting PAD, we have a quantitative metric to measure the quality of synthetic results in terms of age accuracy. Observe that PAD does not compare baseline and follow-up scans. Given that the age predictor is only trained on healthy data it will estimate age on how normal brains will look like. Thus, it should capture when brain ageing acceleration occurs in AD, as others have demonstrated before us [232]. This will increase PAD error when we synthesise with AD or MCI target health state, but given that we use PAD to compare between different methods this error should affect all methods. With advances in brain ageing estimation [241] the fidelity of PAD will also increase. Here since we use PAD to compare across methods even a biased estimator is still a useful method of comparison.

Statistics: All results are obtained on testing sets, and we show average and standard deviation (std, as subscript on all tables), estimated by sample mean and variance on the testing

	SSIM	PSNR	MSE	PAD	CN	MCI	AD
Naïve baseline	$0.71_{0.09}$	$22.1_{3.3}$	$0.097_{0.013}$	$7.2_{3.9}$	$6.3_{3.8}$	$6.8_{3.9}$	$8.7_{4.0}$
Cond. GAN	$0.39_{0.08}$	$14.2_{3.5}$	$0.202_{0.012}$	$9.5_{4.7}$	$8.7_{4.8}$	$9.1_{4.7}$	$10.9_{4.7}$
CycleGAN	$0.46_{0.07}$	$16.3_{3.3}$	$0.193_{0.008}$	$9.7_{5.1}$	$8.9_{4.9}$	$9.4_{5.2}$	$11.0_{5.2}$
CAAE	$0.64_{0.07}$	$20.3_{2.9}$	$0.114_{0.011}$	$5.4_{4.5}$	$4.4_{4.3}$	$5.1_{4.4}$	$6.9_{4.7}$
Ours-previous	$0.73_{0.06}$	$23.3_{2.2}$	$0.081_{0.009}$	$5.0_{3.7}$	$4.0_{3.5}$	$4.6_{3.6}$	$6.6_{4.0}$
Ours	$\mathbf{0.79^{*}_{0.06}}$	$\mathbf{26.1^{*}_{2.6}}$	$\mathbf{0.042^{*}_{0.006}}$	$\mathbf{4.2^{*}_{3.9}}$	$\mathbf{3.1^{*}_{3.6}}$	$\mathbf{3.9^{*}_{3.8}}$	$\mathbf{5.9^{*}_{4.2}}$

Table 5.1: Quantitative evaluation on ADNI dataset (testing set) for several metrics. Columns 2-4 present the results of SSIM, PSNR, MSE, respectively. Columns 5-8 present the overall PAD and the PAD for CN, MCI, and AD data, respectively. We report average and std (as subscript) with **BOLD**, * indicating best performance and statistical significance, respectively (see Section 5.4).

set. We use **bold** font to denote the best performing method (for each metric) and an asterisk (*) to denote statistical significance. We use a paired t-test (at 5% level assessed via permutations) to test the null hypothesis that there is no difference between our methods and the best performing benchmark.

5.5 Results and discussion

We start by showing quantitative and qualitative results on ADNI with detailed evaluation demonstrating quality of the generated images. We then train our model on Cam-CAN to show long-term brain ageing synthesis. We conclude with ablation studies to illustrate the effect of design choices.

5.5.1 Brain ageing synthesis on different health states (ADNI)

In this section, we train and evaluate our model on ADNI dataset, which contains CN, MCI and AD subjects. Our model is trained only on *cross-sectional* data. The results and discussions are detailed below.

5.5.1.1 Quantitative results

The quantitative results are shown in Table 5.1, employing the metrics defined in Section 5.4. For ADNI we also obtained a non-learned naïve baseline that simply calculates performance comparing ground-truth baseline and follow-up images. The naïve baseline result is obtained by subtracting from the followup the baseline (input) image. We involve this non-learned baseline as a lower bound to check if the proposed algorithm synthesises images that are closer to the follow-up than the baseline images or not. As reported in Section 4, the average age prediction error (MAE) of the age predictor on the ADNI testing data is 3.9 years. Estimating PAD separately for CN, MCI and AD testing subjects (see Table 5.1) shows that the best PAD results are obtained on healthy (CN) data. This is expected as the age predictor used to estimate PAD it is trained on healthy data only. However, this bias affects all methods, and thus still allows comparisons between them. Indeed, we can observe that our method achieves the best results in all metrics, with second best being the previous (more simple incarnation) [15] of the proposed model. Embedding health state improves performance, because it permits the method to learn an ageing function specific for each state as opposed to the one learned by the method in [15]. The other benchmarks achieve a lower performance compared to the baseline. The next best results are achieved by CAAE [193], where age is divided into 10 age groups and represented by a one-hot vector. To generate the aged images at the target age (the age of the follow-up studies), we use the age group to which the target age belongs, *i.e.* if the target age is 76, then we choose the age group of age 75-78. We see the benefits of encoding age into ordinal vectors, where the difference between two vectors positively correlates with the difference between two ages in a finely-grained fashion. CycleGAN and Conditional GAN achieve the poorest results unsurprisingly, since conditioning here happens explicitly by training separate models according to different age groups.

5.5.1.2 Qualitative results

Visual examples on two images from ADNI, are shown in Fig. 5.7. For both examples, our method generates most accurate predictions, followed by our previous method [15], offering visual evidence to the observations above. The third best results are achieved by CAAE, where we can see more errors between prediction $\hat{x_o}$ and ground-truth x_o. CycleGAN and

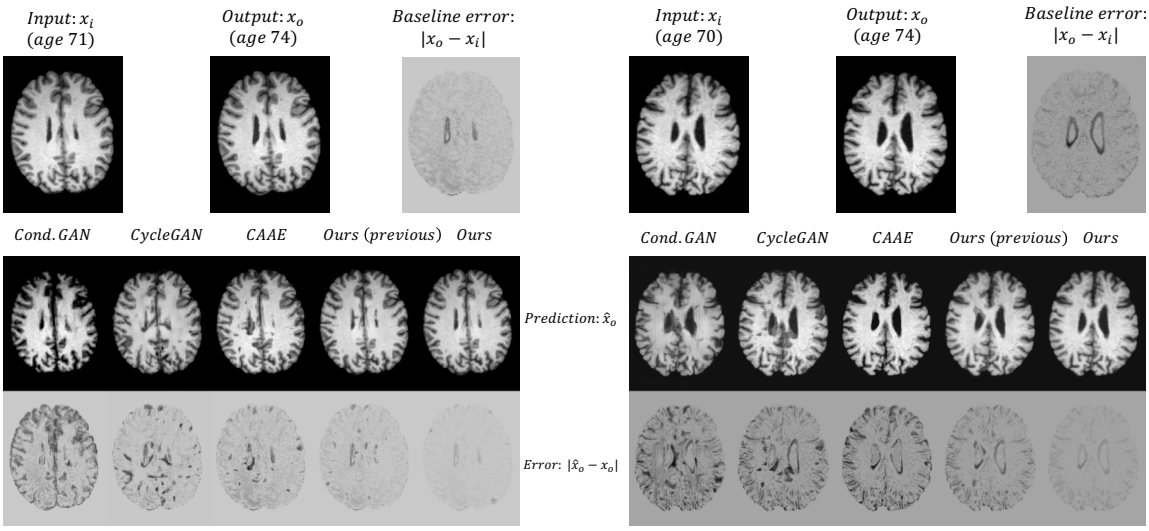


Figure 5.7: Example results of subjects with ground-truth follow-up studies. We predict output $\hat{x}_o$ from input x_i using benchmarks and our method. We also show errors between the outputs and the ground-truths as $|\hat{x}_o - x_o|$. We can observe that our method achieves the most accurate results outperforming our previous method [15] and benchmarks. As a comparison, we also visualized the difference between inputs and ground-truth outputs as $|x_o - x_i|$. For more details see text.

Conditional GAN produced the poorest output images, with observable structural differences from ground-truth, indicating loss of subject identity. We can also observe that the brain ventricle is enlarged in our results and the difference between x_i and x_o is reduced, which is consistent with known knowledge that ventricle increases during ageing.

Furthermore, we show visual results of the same subject at different target health states h_o, in Fig. 5.8. We observe that for all h_o, the brain changes gradually as age (a_o) increases. However, the ageing rate varies based on health state (h_o). Specifically, when h_o is CN, ageing is slower than that of MCI and AD, as one would expect; when h_o is AD, ageing changes accelerate. We also report the estimated ages of these synthetic images as predicted by f_{pred}. While these results show one instance, we synthesised aged images of different health status from 49 ADNI test set CN subjects, with target ages 10 years older than the original age. We then used f_{pred} to estimate the age of these synthetic images. We find that on average, synthetic AD images are 4.9 ± 2.3 years older than the target age, whereas synthetic MCI and CN images are 1.8 ± 2.0 and 1.5 ± 2.1 years older than the target age,

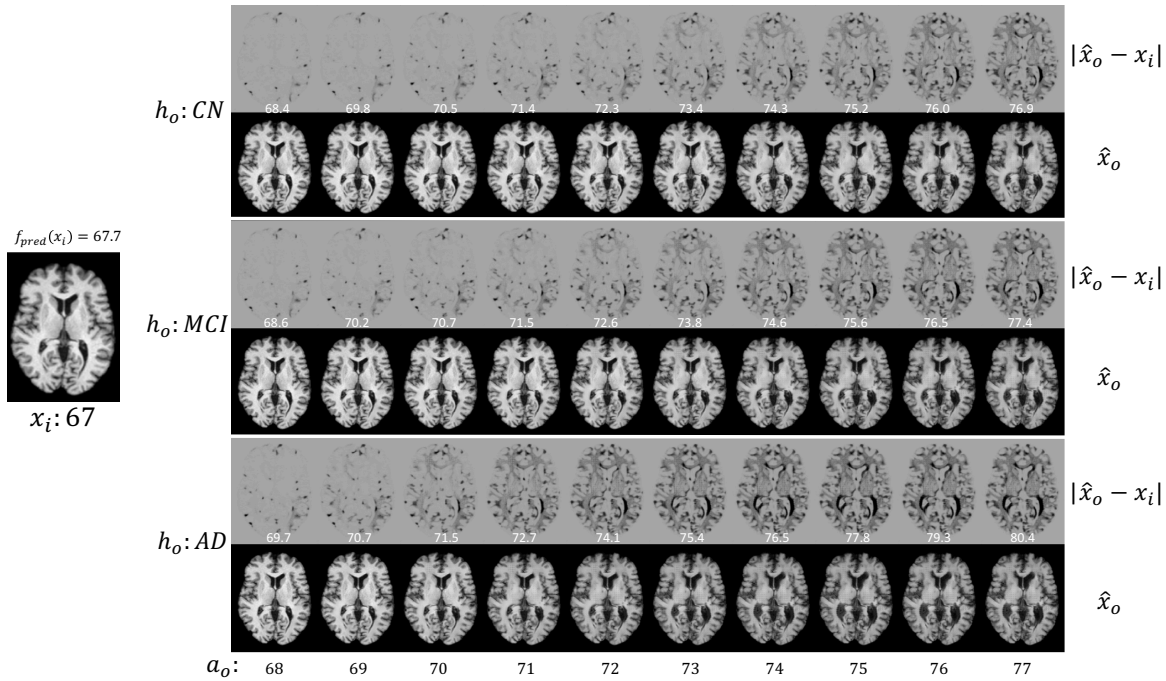


Figure 5.8: Brain ageing progression for a healthy (CN) subject $\mathbf{x}_i$ (at age 67) from ADNI dataset. We synthesise the aged images $\hat{\mathbf{x}}_o$ at different target ages a_o on different health states h_o: CN, MCI and AD, respectively. We also visualise the difference between $\mathbf{x}_i$ and $\hat{\mathbf{x}}_o$, $|\hat{\mathbf{x}}_o - \mathbf{x}_i|$, and show the predicted (apparent) ages of $\hat{\mathbf{x}}_o$ as obtained by our pre-trained age predictor (white text overlaid on each difference image). For more details see text.

respectively. These observations are consistent with prior findings that AD accelerates brain ageing [5]. We also observe that the gray/white matter contrast decreases as age increases, which is consistent with existing findings [244, 245].

5.5.2 Does our model capture realistic morphological changes of ageing and disease?

Here we want to assess whether our model captures known ageing-related brain degeneration. It is known that brain ageing is related to gray matter reduction in middle temporal gyrus (MTG) [246, 57]. We wanted to assess whether synthetic volumes could act as drop-in replacements of ground-truth follow-up in assessing MTG gray matter volume change. We focus here on the MTG as this is well covered by the range of slices we use to train our

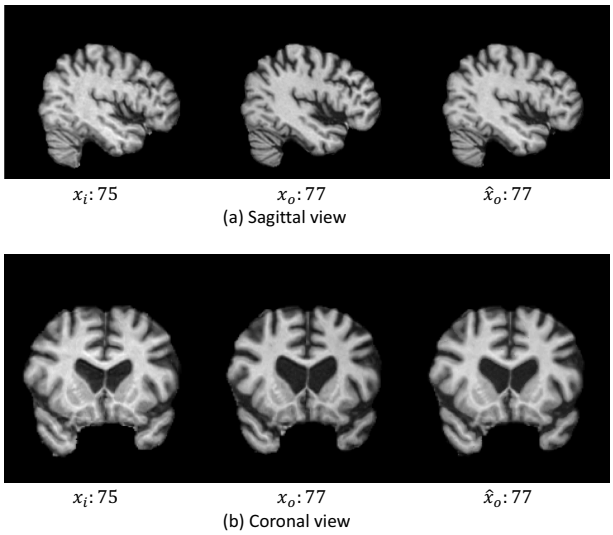


Figure 5.9: Example results of a synthetic 3D volume $\hat{\mathbf{x}}_o$ in sagittal view (top) and coronal view (bottom) from ADNI dataset. Here we construct the 3D volume by stacking the 2D synthetic axial slices of our model. From left to right are slices from a baseline volume $\mathbf{x}_i$, the corresponding follow-up volume $\mathbf{x}_o$, and the stacked synthetic volume $\hat{\mathbf{x}}_o$.

synthesis method. Before we proceed we first illustrate that we can synthesise 3D volumes slice-by-slice, and then show that our model can capture realistic morphological changes.

5.5.2.1 Volume synthesis by stacking 2D slices

We show that, even with our 2D model, we can still produce 3D volumes that show consistency. We applied our model on 2D axial slices and obtained a 3D volume by stacking the synthetic slices. An example result of a stacked synthetic 3D volume in sagittal and coronal views is shown in Figure 5.9. Compared to the respective ground-truth from the same subject, we observe that both sagittal and coronal views of the synthetic volume look realistic and are close to the follow-up images. Note here that our model is trained only on 2D axial slices, for which we chose middle 60 slices from each volume. Our model uses a residual connection and thus makes minimal changes to the regions affected by age instead of synthesising the whole brain image. This helps preserve details and continuity across slices. These results illustrate that we can produce 3D volumes that maintain consistency in different views.

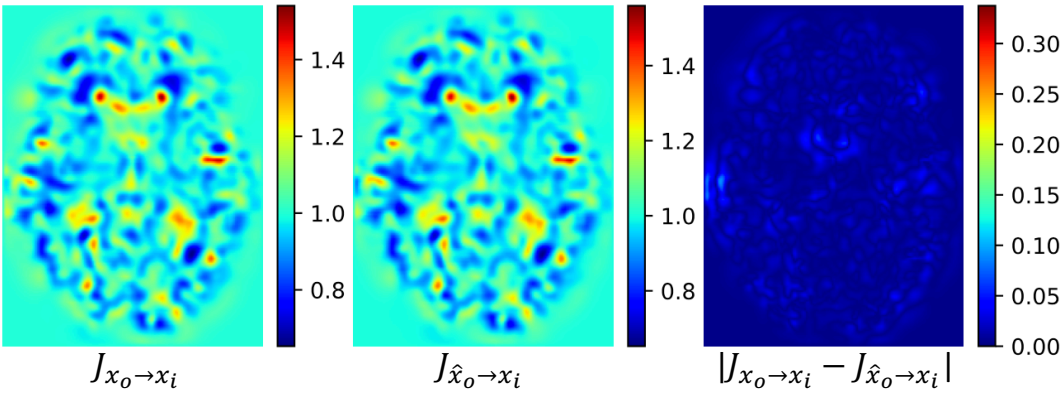


Figure 5.10: An example of Jacobian determinant maps for a subject. From left to right are the Jacobian determinant maps $\mathbf{J}_{\mathbf{x}_o \to \mathbf{x}_i}$, $\mathbf{J}_{\hat{\mathbf{x}}_o \to \mathbf{x}_i}$, and the error map between them: $|\mathbf{J}_{\mathbf{x}_o \to \mathbf{x}_i} - \mathbf{J}_{\hat{\mathbf{x}}_o \to \mathbf{x}_i}|$.

5.5.2.2 Do we capture morphological changes?

We use an ℓ_1 loss to restrict (in pixel space) the amount of change between input and output images. This is computationally efficient, but to show that it also restricts deformations, we measure the deformation between input (baseline) and synthetic or ground-truth follow-up images in ADNI. We obtain for each subject the baseline image $\mathbf{x}_i$, the follow-up image $\mathbf{x}_o$ and the synthetic image $\hat{\mathbf{x}}_o$, respectively. We first rigidly register $\mathbf{x}_o$ to $\mathbf{x}_i$ using Advanced Normalization Tools (ANTs) [247] rigid transformation. Then we non-rigidly register $\mathbf{x}_o$ to $\mathbf{x}_i$ and obtain the Jacobian determinant map $\mathbf{J}_{\mathbf{x}_o \to \mathbf{x}_i}$ that describes the transformation from $\mathbf{x}_o$ to $\mathbf{x}_i$, using ANTs "SyN" transformation [247]. Similarly, we obtain $\mathbf{J}_{\hat{\mathbf{x}}_o \to \mathbf{x}_i}$ that describes the non-linear transformation from $\hat{\mathbf{x}}_o$ to $\mathbf{x}_i$. Fig. 5.10 shows an example of the Jacobian maps for one subject.

From Fig. 5.10, we observe that $\mathbf{J}_{\hat{\mathbf{x}}_o \to \mathbf{x}_i}$ is close to $\mathbf{J}_{\mathbf{x}_o \to \mathbf{x}_i}$. To quantify their difference, we calculate the mean relative error between the Jacobian determinant maps, defined as:

$$E = \mathbb{E}_{\mathbf{x}_i \sim \mathcal{X}_i, \mathbf{x}_o \sim \mathcal{X}_o, \hat{\mathbf{x}}_o \sim \hat{\mathcal{X}}_o} \frac{\|\mathbf{J}_{\mathbf{x}_o \to \mathbf{x}_i} - \mathbf{J}_{\hat{\mathbf{x}}_o \to \mathbf{x}_i}\|_1}{\|\mathbf{J}_{\mathbf{x}_o \to \mathbf{x}_i}\|_1}, \tag{5.6}$$

where $\|.\|_1$ is 1-norm of matrices. We find the mean relative error to be 3.49% on the testing set of 136 images. Similarly, we perform the same evaluations for the results of Conditional GAN, CycleGAN, CAAE and our previous method, and find the mean relative errors to

be 9.87%, 8.76%, 5.91% and 4.43%, respectively. Both qualitative and quantitative results suggest that synthetically aged images capture realistic morphological changes of the brain ageing process.

5.5.2.3 Measuring middle temporal gyrus (MTG) gray matter atrophy.

We further evaluate the quality of the synthetic results by assessing if they can act as a drop-in replacement to real data in a simple study of brain atrophy. We performed ageing synthesis with our model on 136 ADNI testing subjects, such that for each subject we have: a baseline image $\mathbf{x}_i$; a real follow-up image $\mathbf{x}_o$; and a synthetic image $\hat{\mathbf{x}}_o$ of the same target age and health state as of $\mathbf{x}_o$. We then assembled volumes by stacking 2D images. Then we affinely registered both $\mathbf{x}_o$ and $\hat{\mathbf{x}}_o$ and the Human-Brainnetome based on Connectivity Profiles (HCP) atlas [248] to $\mathbf{x}_i$. After that, we obtained the MTG segmentation of $\mathbf{x}_i$, $\mathbf{x}_o$ and $\hat{\mathbf{x}}_o$ by means of label propagation from HCP using the deformation fields. Then we obtained the gray matter segmentation of $\mathbf{x}_i$ using FSL-FAST [249]. The gray matter segmentation of x_o and $\hat{\mathbf{x}}_o$ was subsequently obtained by non-linearly registering $\mathbf{x}_o$ and $\hat{\mathbf{x}}_o$ to $\mathbf{x}_i$ and propagating anatomical labels using ANTs [247]. These steps yield the MTG gray matter volume of $\mathbf{x}_i$, $\mathbf{x}_o$ and $\hat{\mathbf{x}}_o$, termed as $\mathbf{V}_{base}$, $\mathbf{V}_{fol}$, and $\mathbf{V}_{syn}$, respectively. Then, we calculate the relative change between $\mathbf{V}_{base}$ and $\mathbf{V}_{fol}$ as $RC_{real} = \frac{\mathbf{V}_{fol}-\mathbf{V}_{base}}{\mathbf{V}_{base}}$, and the relative change for synthetic data as $RC_{syn} = \frac{\mathbf{V}_{syn}-\mathbf{V}_{base}}{\mathbf{V}_{base}}$. We repeat this for several subjects in three patient type groups, *i.e.* CN (49), MCI (46) and AD (41).

We expect, following [246] and [57], to see a statistical relationship between patient type and RC_{real} when assessed with a one-way analysis of variance (ANOVA). If a similar relationship is shown also with synthetic data RC_{syn}, it will demonstrate that for this statistical test, our synthetic data can act as a drop-in replacement to real data, and as such have high quality and fidelity.

The results are summarised in Table 5.2, where we report also the F-statistic of the omnibus one-way ANOVA test. We observe that MTG gray matter volume reduces in both real and synthetic volumes. This indicates that our synthetic results achieve good quality and similar statistical conclusions can be drawn using real or synthetic data in this simple atrophy study.

	Relative change	F-statistic
real (RC_{real})	$-0.071_{\pm 0.0096}$	4.008^*
synthetic (RC_{syn})	$-0.083_{\pm 0.0099}$	4.539^*

Table 5.2: Analysis of MTG gray matter relative change between baseline and follow-up real or synthetic. Mean and std are reported as well as the corresponding F-statistic of a one-way ANOVA test (between relative change and patient type), with asterisk indicating significance ($p < 0.05$).

5.5.3 Long term brain ageing synthesis

In this section, we want to see how our model performs in long term brain ageing synthesis. As ADNI dataset only covers old subjects, we use Cam-CAN dataset which contains subjects of all ages. We train our model with Cam-CAN dataset where no longitudinal data are available, but evaluate it on the longitudinal part of ADNI to assess the generalisation performance of our model when trained on one dataset and tested on another.

5.5.3.1 Qualitative results

In Fig. 5.11, we demonstrate the simulated brain ageing process throughout the whole lifespan, where the input images are two young subjects from Cam-CAN. We observe that the output gradually changes as a_o increases, with ventricular enlargement and brain tissue reduction. This pattern is consistent with previous studies [250, 251], showing that our method learns to synthesise the ageing brain throughout the lifespan even trained on cross-sectional data.[10] Fig. 5.11 offers only a qualitative visualization to show the potential of life-time simulation. We cannot quantitatively evaluate the quality of these synthetic images due to the lack of longitudinal data in Cam-CAN. However, both the previous section on ADNI where we train and test on ADNI, and the next section, where we use longitudinal ADNI as testing set we but train on Cam-CAN data, offer considerable quantitative experiments.

[10]We observe checkerboard artefacts near the ventricles after target age 67. Such artefacts are a known problem in computer vision and mostly likely due to the use of deconvolutional layers in the decoder [252].

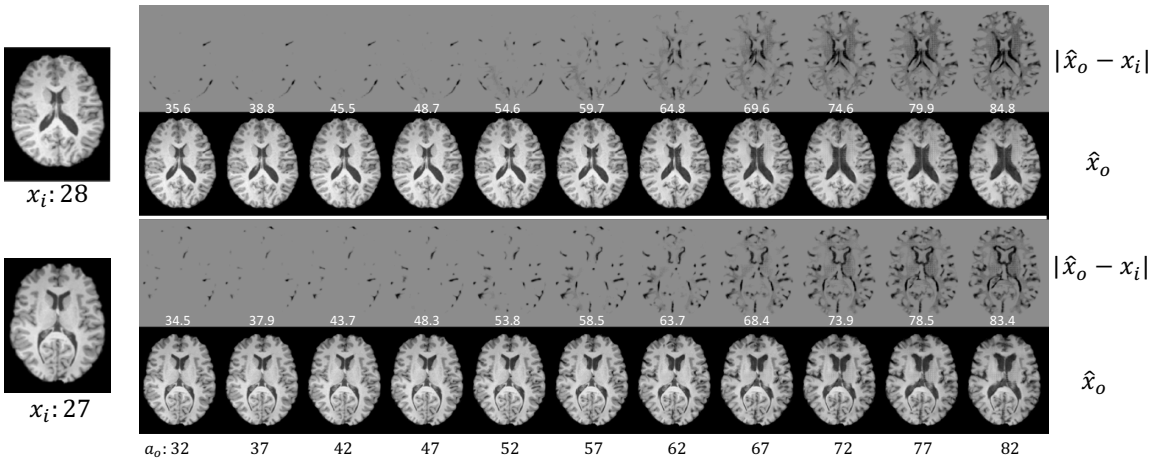


Figure 5.11: Long-term brain ageing synthesis on Cam-CAN dataset. We synthesise the aged images $\hat{\mathbf{x}}_o$ at different target ages a_o and show the difference between input images $\mathbf{x}_i$ and $\hat{\mathbf{x}}_o$, $|\hat{\mathbf{x}}_o - \mathbf{x}_i|$, and show the predicted (apparent) ages of $\hat{\mathbf{x}}_o$ as obtained by our pre-trained age predictor (white text overlaid on each difference image). Note here $\mathbf{x}_i$: N means an input image at age N. For more details see text.

5.5.3.2 Quantitative results (generalisation performance on ADNI)

To evaluate how accurate our longitudinal estimation is, even when training with cross sectional data from *another* dataset, we train a model on Cam-CAN and evaluate it on ADNI. We use the longitudinal portion of ADNI data, and specifically only the CN cohort, to demonstrate generalisation performance.[11] Given an image of ADNI we use our Cam-CAN trained model to predict an output at the same age as the real follow up image. We compare our prediction with the ground truth follow up image (in the ADNI dataset). The results are shown in Table 5.3. We observe that though our model is trained and evaluated on different datasets, it still achieves comparable results with those of Table 5.1 and outperforms benchmarks.

[11]We purposely do not use any intensity harmonisation that uses both datasets, *e.g.* histogram matching. Such methods will leak information from ADNI to Cam-CAN. Any leakage would skew (to our favour) the generalisation ability which we want to avoid. Thus, our experiments also indirectly evaluate how design choices (*e.g.* using a residual connection in the generator) help with differences in intensities between datasets.

	SSIM	PSNR	MSE	PAD
Cond. GAN	$0.38_{\pm 0.12}$	$13.9_{\pm 4.2}$	$0.221_{\pm 0.021}$	$11.3_{\pm 5.6}$
CycleGAN	$0.42_{\pm 0.09}$	$14.4_{\pm 3.8}$	$0.212_{\pm 0.016}$	$10.2_{\pm 5.5}$
CAAE	$0.59_{\pm 0.10}$	$19.3_{\pm 3.9}$	$0.121_{\pm 0.012}$	$5.9_{\pm 4.7}$
Ours-previous	$0.68_{\pm 0.08}$	$22.7_{\pm 2.8}$	$0.095_{\pm 0.014}$	$5.3_{\pm 3.8}$
Ours	$\mathbf{0.74}^{*}_{\pm \mathbf{0.08}}$	$\mathbf{24.2}^{*}_{\pm \mathbf{2.7}}$	$\mathbf{0.043}^{*}_{\pm \mathbf{0.009}}$	$\mathbf{5.0}_{\pm \mathbf{3.6}}$

Table 5.3: Quantitative evaluation of methods trained on Cam-CAN and evaluated on ADNI.

	SSIM	PSNR	MSE
L_{GAN}	$0.55_{\pm 0.14}$	$18.4_{\pm 3.7}$	$0.132_{\pm 0.013}$
$L_{GAN} + L_{rec}$	$0.62_{\pm 0.12}$	$19.6_{\pm 3.2}$	$0.089_{\pm 0.014}$
$L_{GAN} + L_{ID}$	$0.74_{\pm 0.07}$	$24.3_{\pm 2.5}$	$0.074_{\pm 0.010}$
$L_{GAN} + L_{ID} + L_{rec}$	$\mathbf{0.79}^{*}_{\pm \mathbf{0.08}}$	$\mathbf{26.1}^{*}_{\pm \mathbf{2.6}}$	$\mathbf{0.042}^{*}_{\pm \mathbf{0.006}}$

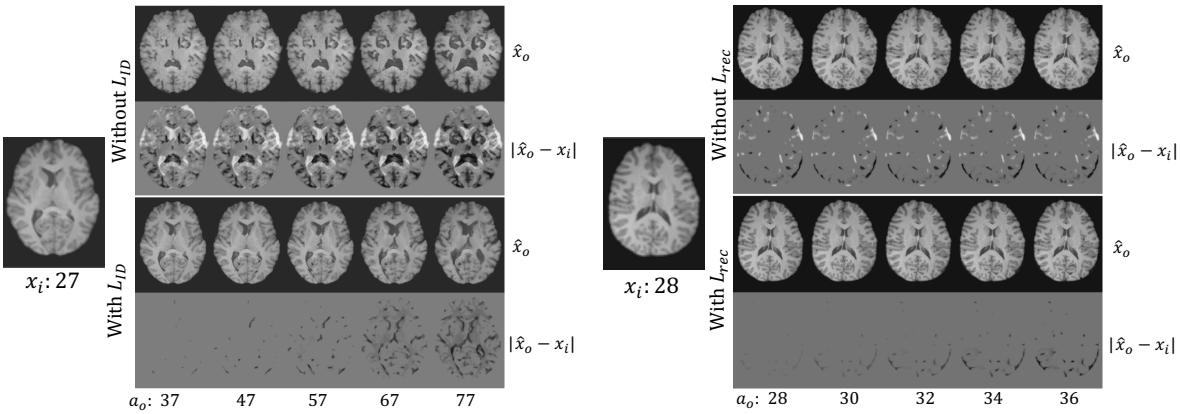
Table 5.4: Ablations on using different combinations of cost functions.

5.5.4 Ablation studies

We ablate loss components, explore different conditioning mechanisms, and explore latent space dimensions.

5.5.4.1 Effect of loss components

We demonstrate the effect of L_{ID} and L_{rec} by assessing the model performance when each component is removed. In Table 5.4 we show quantitative results on ADNI dataset. In Fig. 5.12 we illustrate qualitative results on Cam-CAN dataset to visualise the effect. We can observe that the best results are achieved when all loss components are used. Specifically, without L_{ID}, the synthetic images lost subject identity severely throughout the whole progression, *i.e.* the output image appears to come from a different subject; without L_{rec}, output images suffer from sudden changes at the beginning of progression, even when $a_o = a_i$. Both quantitative and qualitative results show that the design of L_{ID} and L_{rec} improves preservation of subject identity and enables more accurate brain ageing simulation.



(a) Ablation study on *identity loss* L_{id} (b) Ablation study on *self-reconstruction loss* L_{rec}

Figure 5.12: Ablation studies for loss components. **Left:** ablation study of L_{ID}. Top row shows that without L_{ID}, the network can lose the subject identity. Bottom row shows that the use of L_{ID} can enforce the preservation of subject identity, such that the changes as ages are smooth and consistent. **Right:** ablation study on L_{rec}. When L_{rec} is not used (top two rows), there are sudden changes at the beginning of ageing progression simulation (even at the original age), which hinders the preservation of subject identity. In contrast, when L_{rec} is used (bottom two rows), the ageing progression is smoother, which demonstrates better identity preservation. Note here x_i: N means an input image at age N.

5.5.4.2 Effect of different embedding mechanisms

We investigate the effect of different embedding mechanisms. Our embedding mechanism is described in Section 5.3. We considered to encode age as a normalized *continuous* value (between 0 and 1) or using a *one-hot* vector, which was then concatenated with the latent vector at the bottleneck. The qualitative results are shown in Fig. 5.13. We can see that when age is represented as a normalized *continuous* value, this is ignored by the network, thus generating similar images regardless of changes in target age a_o. When we use *one-hot* vectors to encode age, the network still generates realistic images, but the ageing progression is not consistent, *i.e.* synthetic brains appear to have ventricle enlarging or shrinking in random fashion across age. In contrast, with *ordinal encoding*, the model simulates the ageing process consistently. This observation is confirmed by the estimated ages of the output images by f_{pred}.

We also compare with an embedding strategy where we concatenate $\mathbf{h}_o$, $\mathbf{a}_d$ and the bottleneck latent vector $\mathbf{v}_1$ together, and the concatenated vector is processed by the Decoder to generate

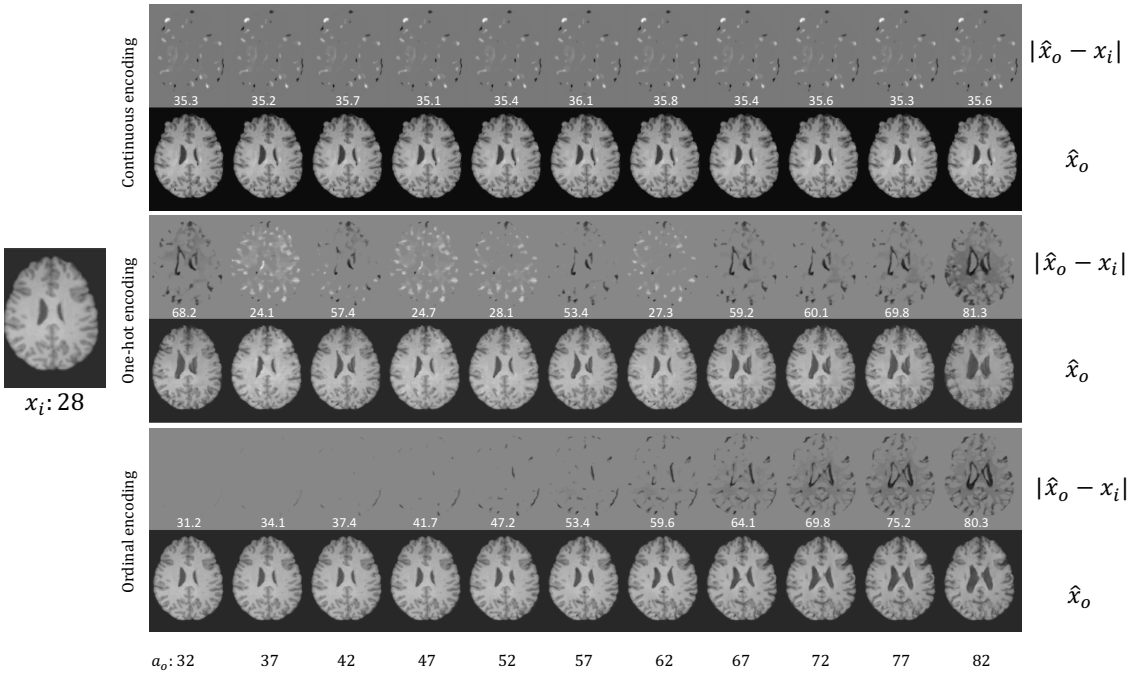


Figure 5.13: Example results for *continuous*, *one-hot* and *ordinal* encoding on the Cam-CAN dataset for an image ($\mathbf{x}_i$) of a 28 year old subject. We synthesise aged images $\hat{\mathbf{x}}_o$ at different target ages a_o. We also show the difference between $\mathbf{x}_i$ and $\hat{\mathbf{x}}_o$, $|\hat{\mathbf{x}}_o - \mathbf{x_i}|$, and report estimated age (white text overlaid at the bottom of each difference image). The proposed ordinal encoding shows consistent and progressive changes.

the output image. We refer to this embedding strategy as $concat_{all}$. Results on ADNI are shown in Table 5.5. We found with $concat_{all}$, the network tends to ignore the health state vector $\mathbf{h}_o$ and only use $\mathbf{a}_d$. This can be caused by the dimensional imbalance between $\mathbf{h}_o$ (2×1) and $\mathbf{a}_d$ (100×1). When *one-hot encoding* is used, performance deteriorates even more.

	SSIM	PSNR	MSE	PAD
One-hot	$0.54_{\pm 0.14}$	$17.3_{\pm 3.8}$	$0.177_{\pm 0.014}$	$9.7_{\pm 4.9}$
concat$_{\text{all}}$	$0.74_{\pm 0.09}$	$23.9_{\pm 2.9}$	$0.065_{\pm 0.011}$	$5.2_{\pm 3.9}$
Ours	$\mathbf{0.79}^{*}_{\pm \mathbf{0.08}}$	$\mathbf{26.1}^{*}_{\pm \mathbf{2.6}}$	$\mathbf{0.042}^{*}_{\pm \mathbf{0.006}}$	$\mathbf{5.0}_{\pm \mathbf{3.6}}$

Table 5.5: Quantitative results of different embedding mechanisms.

	SSIM	PSNR	MSE	PAD
65×1	$0.73_{\pm 0.09}$	$23.6_{\pm 3.1}$	$0.065_{\pm 0.012}$	$5.6_{\pm 4.1}$
260×1	$0.76_{\pm 0.10}$	$24.9_{\pm 2.9}$	$0.055_{\pm 0.012}$	$5.3_{\pm 3.8}$
130×1 (ours)	$\mathbf{0.79}^{*}_{\pm \mathbf{0.08}}$	$\mathbf{26.1}^{*}_{\pm \mathbf{2.6}}$	$\mathbf{0.042}^{*}_{\pm \mathbf{0.006}}$	$\mathbf{5.0}_{\pm \mathbf{3.6}}$

Table 5.6: Quantitative results of different choices of the v_2 dimension.

5.5.4.3 Effect of latent space dimension

We explored whether latent dimension affects performance. We altered the length of the latent vector ($\mathbf{v}_2$) from 130×1 to twice smaller/larger and compared the corresponding models on ADNI. Our findings are shown in Table 5.6. We find that our choice (130×1) achieved the best results. It appears that too small is not enough to represent image information well, and too large can cause dimension imbalance.

5.5.4.4 Comparison with longitudinal model

To compare our method with models that use longitudinal data , we create a new benchmark where we train a fully supervised generator using only longitudinal ADNI data. The results are shown in Table 5.7. We see that our method has slightly better performance than the longitudinal model. This is because the proposed model is trained on 786 subjects (cross-sectional data), while the longitudinal model is trained on a longitudinal cohort of ADNI of 98 subjects. This illustrates the benefit of using cross-sectional data. Note that our SSIM results are similar to those presented in [191].

	SSIM	PSNR	MSE
Longitudinal	$0.72_{\pm 0.09}$	$24.2_{\pm 3.0}$	$0.076_{\pm 0.013}$
Ours	$0.79_{\pm 0.08}$	$26.1_{\pm 2.6}$	$0.042_{\pm 0.006}$

Table 5.7: Quantitative results of a longitudinal benchmark and our method.

5.5.4.5 Data augmentation for AD classification

We explore whether we can use our model to generate synthetic data used to augment training sets for training an Alzheimer's disease classifier. We select 200 ADNI subjects as training data (100 AD, 100 CN), 40 subjects as validation data (20 AD, 20 CN), and 80 (40 AD, 40 CN) subjects as testing data. For each subject, there are 60 2D slices. Next, we train classifiers of the same VGG architecture to classify AD and CN subjects varying the composition of the training data combining real and synthetic data obtained by our generator. We always evaluate the classifiers on the same testing set. The synthetic data are generated from the training set using our proposed method by randomly selecting target ages larger than the original ages. As shown in Table 5.8, we first train classifiers only on real data varying the size of the training data (1st and 2nd rows). Then we compose mixed sets of the same size of 200 subjects varying the ratio of real vs. synthetic data (3rd and 4th rows), *e.g.* 10%+90% means this set is composed of 10% real data and 90% synthetic data. Note here the 90% synthetic data are not generated from the whole training set, but from the 10% real data.

We can observe that when training on 10% of real training data, the accuracy reduces by almost 40% compared to when using the full training data. However, the performances improve when synthetic data are involved. The results demonstrate that our method can be used as data augmentation to improve AD classification especially when the training data are not sufficient.

Real data	10%	30%	50%	70%	100%
Accuracy (%)	51.3	55.7	64.6	74.0	89.5
Real data + synthetic data	10%+90%	30%+70%	50%+50%	70%+30%	
Accuracy (%)	58.7	64.0	72.6	81.7	

Table 5.8: Quantitative results of VGG-based AD/CN classifiers trained on different datasets. The first two rows show results when trained on varying size of real training data, *e.g.* 10% means this model is trained on 10% of the real training data; the last two rows show results when trained on mixed datasets with different ratios of real and synthetic data, *e.g.* 10%+90% means this model is trained on 10% real training data and 90% synthetic data.

Furthermore, we perform another experiment to demonstrate our model's potential to improve the classification accuracy for specific age groups and thus target augmentation to treat data imbalance. We evaluate the classification model trained with 100% real data on test set subjects of age 65 to 70 years old. We find an accuracy of 67.2%, which is much lower than the overall accuracy (89.5%, Table 5.8). This may be likely due to training data imbalance: we have only 5 training subjects with age between 65 and 70 yrs. To alleviate this data imbalance, we use our model to generate 25 synthetic subjects with target ages between 65 and 70 yrs from younger subjects in the training set. Then we train a new AD classifier on 100% real data and the 25 synthetic subjects, and evaluate its performance on the same testing and age group. Accuracy now increases to 80.1% a substantial change from 67.2%.

5.6 Summary

We present a method that learns to simulate subject-specific aged images *without* longitudinal data. It relies on a Generator to generate the images and a Discriminator that captures the joint distribution of brain images and clinical variables, *i.e.* age and health state (AD status). We propose an embedding mechanism to encode the information of age and health state into our network, and age-modulated and self-reconstruction losses to preserve *subject identity*. We present qualitative results showing that our method is able to generate consistent and realistic images conditioned on the target age and health state. We evaluate with longitudinal data from ADNI for image quality and *age accuracy*. We demonstrate on ADNI and Cam-CAN datasets that our model outperforms benchmarks both qualitatively and quantitatively and, via a series of ablations, illustrate the importance of each design decision. The reliability of the generated images is measured with quantitative metrics, including a drop-in replacement to real data in a simple study of brain atrophy. However, the golden standard of reliability should be the utility in practical clinical applications. We use deterministic models to simulate brain ageing trajectory for one subject given certain conditional factors. However, in future work, we could consider using probabilistic models that can predict different ageing trajectories with probability. This will increase the variability of the generated images. In the next Chapter, a simple procedure will be proposed to utilise the generative models for downstream tasks.

Chapter 6
Utilising Pre-trained Generative Models for Downstream Tasks

6.1 Introduction

In previous chapters, we focused on developing deep generative models that can synthesise realistic medical images, conditioned on the existence of pathology or age. In this chapter, we focus on another direction, *i.e.* how to utilise generative models for downstream tasks. Specifically, we choose the classification of Alzheimer's Disease as the downstream task and the brain ageing generative model (proposed in Chapter 5) as the generative model. We propose a procedure to utilise the generative model to improve AD classification performance via adversarial training between a conditional factor (*i.e.* target age) and the AD classifier.

Deep learning heavily relies on the large size and quality of training data. In some cases where only limited training data are available, deep neural networks tend to memorise the data and cannot generalise well to unseen data [253, 254]. This is known as *over-fitting* [253]. One of the most popular approaches to overcome over-fitting is *data augmentation* [255].

Conventional data augmentation approaches mainly apply random image transformations, such as cropping, horizontal mirroring, rotation, and intensity transformations. However, one problem with these conventional data augmentations is that augmentation that works well for one dataset may not transfer well to another [256]. Furthermore, as mentioned in [257] traditional data augmentation methods may introduce *distribution shift*, *i.e.* the joint distribution of inputs and outputs changes, and consequently hurt the performance on unaugmented data during inference.

Some recent works aimed to solve this problem by learning parameters for data augmentation that can better improve the downstream task performances [258, 259, 256, 260, 261, 262]. However, these approaches are still based on traditional image transformations, *e.g.* cropping,

rotation and deformation, but do not consider semantic transformation [263]. For instance, say we have an image of a person who wears glasses, and the task is to predict his/her gender, instead of performing rotations or mirroring, we augment this data by making an image of the same person who does not wear glasses. This way of data augmentation could be considered as complementary to traditional techniques.

To achieve the augmentation mentioned above, one way to augment the training data is to train a deep generative model and use the synthetic output of the generative model as augmentations [264, 265, 266, 267, 265, 268]. However, these approaches focus more on the training stage of generative models and randomly generate samples for data augmentation, but did not consider how to utilise these models. Some recent works [269, 270] proposed to first parse parts of human body [269] or face [270] resulting in a *part pool*, and then create *hard* samples by pasting patches from the part pool to images in an adversarial way. However, parsing and composing parts of body or organs is hard for medical data and can affect key clinical information specific to each subject.

Furthermore, a recent work [258] proposed implicit semantic data augmentation (ISDA) to augment the data in the latent space. This method has been extended to alleviate *long-tailed* distribution by adopting meta learning [271, 272]. However, ISDA first estimated the covariance matrix of features for each *class*, and sampled *directions* from normal distributions with these covariance matrices. Then they translated along these directions to augment data. However, they focused on natural image datasets where images of different classes (*e.g.* car, dog, cloud, *etc.*) are easily distinguishable, and thus the class-matrix covariance matrices are easy to estimate. For the task of AD diagnosis, the brain images are very similar to each other, with only subtle differences. It could be hard to estimate the class-wise covariance matrices. More importantly, there is no loss or metrics to guarantee that the found *directions* are meaningful for the brain.

Here we propose a simple procedure to utilise a pre-trained generative model to improve the performance of a downstream classifier. The proposed procedure formulates an adversarial game between the conditional input of the generator and the classifier. The key idea is to find which *conditional factor* can result in the *hard* counterfactuals (synthetic output of the generator) for the classifier, which can be viewed as finding the 'weakness' of the classifier.

Then we force the classifier to overcome its 'weakness' by training it on these *hard* synthetic samples.

In this chapter, we choose the classification of Alzheimer's Disease as the downstream task and utilise a pre-trained brain ageing synthesis model to improve the AD classifier. We conduct a series of experiments to show the effectiveness of our method. We first show that the proposed approach can improve the test accuracy and Area Under the ROC Curve (AUC) of a pre-trained AD classifier (Section 6.5.1). Then we consider the usage of generative models in a *continual learning* context and show the proposed method can alleviate *catastrophic forgetting* (Section 6.5.2). Furthermore, we show the proposed approach can be used to alleviate *spurious correlations* (Section 6.5.3). Finally, we provide an ablation study to explain why the adversarial game is between the target age a and the classifier C, instead of C and the generator G (Section 6.5.4).

Our main contributions are as follows:

- We propose a simple procedure to utilise a pre-trained brain ageing generation model for a downstream AD classifier, with an adversarial game between the *target age* and the *classifier*.
- We consider the scenario of using generative models in a *continual learning* context and show that our approach can help alleviate *catastrophic forgetting*.
- We conduct a simple experiment to show that the proposed approach has the potential to alleviate *spurious correlations*.
- We provide an ablation study to explain the choice of not updating the generator in the adversarial game.

The chapter proceeds as follows: Section 6.2 reviews related work on data augmentation. Section 6.3 details the proposed procedure. Section 6.4 describes the experimental setups. Section 6.5 presents results and discussion. Finally, Section 6.6 concludes the chapter with limitations and future directions.

6.2 Related works

Data augmentation is an important technique that aims to increase the number of available training data for deep learning models [255]. Earlier research on data augmentation focused on alleviating over-fitting [101]. However, apart from over-fitting, deep learning models could suffer from other data issues. For example, *domain shift* happens when training data are drawn from a probability distribution that is different from that of practical data, which will hinder the practical use of deep models [273]. Furthermore, data imbalance can result in *long-tailed distribution*, i.e. several majority classes contain most of the samples while other minority classes only contain a few samples. Moreover, *spurious correlations* occur when the two factors appear to be correlated, but in fact, they are not, which is another consequence of data imbalance [274]. To solve these issues, more advanced data augmentation approaches are required.

Conventional Data Augmentation: Conventional techniques to augment data include rotation, scaling, intensity manipulations, *etc.* These methods are simple and often effective. For instance, Shina et al. [275] proposed Negative Data Augmentation (NDA) which intentionally destroyed the global spatial coherence of images. The resulting images were used as negative data to improve the performance of GANs by training to avoid the negative distribution. However, this approach is targeted to improve GAN training, but may not be suitable for other tasks. For ImageNet [21], the augmentation approach proposed in [101] still remains the standard. One problem with these conventional data augmentations is that augmentations that work well for one dataset may not transfer well to another [256]. Furthermore, as pointed by [257], traditional data augmentation methods may hurt the performance on unaugmented data during inference.

Reinforcement Learning (RL) based Data Augmentation: RL has been used to find the optimal policies to perform data augmentation. AutoAugment is proposed in [256] to learn the optimal policies to augment data. These policies contain the choice of the image operations, *e.g.* translation, rotation or shearing, and the probabilities and magnitudes with which the functions are applied. Adversarial AutoAugment [276] extended [256] by changing the reward game to an adversary, *i.e.* finding policies that increase the target network loss. Sim-

ilarly, a recent work [277] proposed to select and compose pre-specified base data transformations (such as rotations, shears, central swirls for images) into a more sophisticated "tool chain" for data augmentation. Different from the methods focused on finding parameters for image transformation, [278] used RL to select data from synthetic data pool to improve downstream tasks. Nevertheless, a common problem for these approaches is instability and difficulty of the training of RL [279], which limits their utility.

Adversarial Data Augmentation with Conventional Techniques: These approaches focus on finding augmentation parameters that can better improve the downstream tasks in an adversarial way. For instance, in [261] the authors augmented data using virtual adversarial training (VAT) [280] which adds addictive adversarial noise, while [259] performed adversarial training with a chain of image manipulations, *e.g.* scaling, rotation, translation, *etc.* Similarly, a recent work[260] learned to augment data by training three generators, each controls one image operation: affine translation, deformation and additive noise masks. Furthermore, a GAN-based data augmentation method was proposed in [281] to improve cardiac image segmentation, where a deformation field generator and an intensity field generator were trained with adversarial loss and supervised loss to modify data and labels. However, these methods are still based on traditional image transformations, *e.g.* rotation and affine translation, and cannot augment the *semantic information*. For instance, consider the task of AD diagnosis (classification), rotation, scaling and cropping do not change the AD diagnosis and age of brains while deformation will affect the clinical information of brains.

Data Augmentation using Generative Models: A more direct way to augment data semantically is to use generative models. For example, PGGAN was used in [268] to generate pairs of images and segmentations that were used for data augmentation. However, there was no loss to enforce that the synthetic image and segmentations are paired. Data Augmentation Generative Adversarial Network (DAGAN) was proposed in [266], with an encoder to extract features from images and a decoder to produce synthetic data taking these features plus random noise as input. This method required pairs of images for augmentation, which limits its utility. Balancing GAN (BAGAN) was proposed in [267] to alleviate the difficulty of training GANs with an imbalanced dataset by learning useful features from majority classes and using these features to generate minority classes. Similarly, in [265] a GAN-based approach was

proposed to restore balance in an imbalanced dataset by incorporating the majority distribution structure in the generation of new minority samples. Furthermore, a recent work [264] proposed a class-conditional GAN that improved classification accuracy in low data regimes. However, these approaches focused on how to train a generator but did not consider how to utilise this generator. By contrast, our approach can guide the generator to produce synthetic data that is *useful* for improving downstream tasks via an adversarial training scheme.

Data Augmentation via counterfactuals: Pearl's ladder of causation defines three levels of causal hierarchy [282]. Specifically, the first level of the causation is *association* that deals with questions of the type "*What are the data that I observe?*", the second level is *Intervention* that concerns questions such as "*What will happen if I do A?* ", and the third level is *counterfactual* that aims to answer "*What would have happened if I had done A instead of B?*" [282, 196]. Following this definition, the proposed models in Chapter 4 and 5 are, in nature, counterfactual generators, as we try to answer questions such as "*What would a brain look like if it had not got a tumour?* "and "*What would the brain of a subject look like if he or she were at the age X?* "As such, here we review related works that use counterfactuals for data augmentation.

Some recent work [283, 284] focused on non-image data where each data point is composed of several feature values, *e.g.* breed, age and milk-yield of a cow, and generated counterfactuals by simply transferring some feature values from one data sample to another. While these methods achieved good performance, they could not be applied to image data as changing features of images is a more challenging task. To generate counterfactuals for images, a recent work [285] localised the foreground and background of an image and then infilled the foreground or background with artefacts or a GAN. Similarly, Sauer et al. [286] produced counterfactuals by changing the texture of foreground and background of an image with GANs. However, these approaches may not be suitable for medical images where the background is less various than natural images, *e.g.* the background of a typical T1 MRI brain image is black. Moreover, Goel et al. [287] used CycleGANs to transform between different subgroups within a class to alleviate data imbalance and spurious correlation, but they needed to train a CycleGAN for each pair of subgroups, which could be expensive. Furthermore, StylEx [288] tried to explain the decision of a classifier by generating counterfactuals. This

method focused on generating counterfactuals that can explain the classifier, while we focused on improving a classifier by finding *hard* counterfactuals. Other works [289, 290, 291] followed the same scheme: randomly generating counterfactuals by deep generative models and then improving downstream tasks on these counterfactuals, but did not consider which counterfactuals to generate. By contrast, we propose an adversarial training framework to find *hard* counterfactuals that are more helpful for training.

6.3 Methodology

6.3.1 Problem overview

We denote an image as $\mathbf{x} \sim \mathcal{X}$, and a conditional generative model G that takes an image $\mathbf{x}$ and a conditional attribute vector a and generates a counterfactual $\hat{\mathbf{x}}$ that corresponds to a: $\hat{\mathbf{x}} = G(\mathbf{x}, a)$. For each $\mathbf{x}$, there exists a label $y \sim \mathcal{Y}$. We define a downstream classifier C that takes $\mathbf{x}$ as input and predicts the label $\hat{y}$: $\hat{y} = C(\mathbf{x})$. Note here the label y does not need to equal to the conditional factor a. For instance, if $\mathbf{x}$ is a face image of a person, then v could be his/her age, and y could be his/her gender.

Suppose we have a pre-trained generative model G and a pre-trained classifier C. One question is whether we can use the counterfactuals generated by G to improve the performance of C and how to utilise them. Our solution is to find the counterfactuals that are the *most helpful* to improve C. Inspired by the mini-max mechanism in GANs [107] and adversarial learning [292, 280], we formulate an adversarial game between the conditional attributes a and the model C. That is, try to find the conditional attributes that result in *hard* counterfactuals for C, and improve C on these hard samples. This strategy can viewed as finding the 'weakness' of the model C and purposely forcing C to overcome its 'weakness'.

In this chapter, $\mathbf{x}$ is a brain image, y is the AD diagnosis of $\mathbf{x}$, and a represents the target age on which the generator G was conditioned.[1] We use the proposed brain ageing generation

[1]The model G can generate synthetically aged brain images conditioned on age a and health state. Since this chapter does not change the health state (AD state) for the synthetic images, we simplify the definition by only listing age a as the conditional factor.

model proposed in Chapter 5 as G, and a VGG-based [293] AD classification model as C.

6.3.2 Fourier encoding for conditional factors

The brain ageing model proposed in Chapter 5 used *ordinal encoding* to encode the conditional age and health state vectors (see Section 5.3.2). However, in this chapter, as we need to back-propagate gradients to *update* the input conditional vectors, *ordinal encoding* could cause issues, because the encoded vectors are, in nature, discrete and need to maintain a certain shape (with 0 elements always on top of 1).

To enable gradient backpropagation to update the conditional vectors, here we use *Fourier encoding* [294, 295, 296] to encode the conditional attributes, *i.e.* age and heath state (diagnosis of AD). The key idea of Fourier encoding is to map low-dimensional vectors to a higher dimensional domain using a set of sinusoids. For instance, if we have a d-dimensional vector which is normalised into $[0, 1)$, $\mathbf{v} \in [0, 1)^d$, then the encoded vector can be represented as [295]:

$$\gamma(\mathbf{v}) = [p_1 \cos(2\pi \mathbf{b_1^T v}), p_1 \sin(2\pi \mathbf{b_1^T v}), ..., p_m \cos(2\pi \mathbf{b_m^T v}), p_m \sin(2\pi \mathbf{b_m^T v})], \tag{6.1}$$

where $\mathbf{b_j}$ can be viewed as the Fourier basis frequencies, and p_j^2 the Fourier series coefficients.

In this chapter, the vector $\mathbf{v}$ represents the target age a and the health status (AD diagnosis), and $d = 2$. In practice, we set $p_j^2 = 1$ for $j = 1, ..., m$, and $\mathbf{b_j}$ are independently and randomly sampled from a Gaussian distribution, $\mathbf{b_j} \sim \mathcal{N}(\mu_{scale} * \mathbf{I}, 0)$, where μ_{scale} is a hyperparameter and set to 10 in this chapter. We set $m = 100$ and the resulting $\gamma(\mathbf{v})$ is 200d.

Fourier encoding has been experimentally shown to effectively represent a low-dimensional signal in a way that neural networks can capture [295, 296]. In our experiment, we also found that after using Fourier encoding, the brain ageing network achieved similar results qualitatively and quantitatively as the original model in Chapter 5. Table 6.1 presents the quantitative results of brain models using *ordinal encoding* and *Fourier encoding*. The quantitative evalu-

ation is performed in the same way as in Section 5.5.1 and Table 5.1. From Table 6.1 we can observe that *Fourier encoding* achieves very similar quantitative results as *ordinal encoding*, demonstrating its effectiveness to encode age and health status.

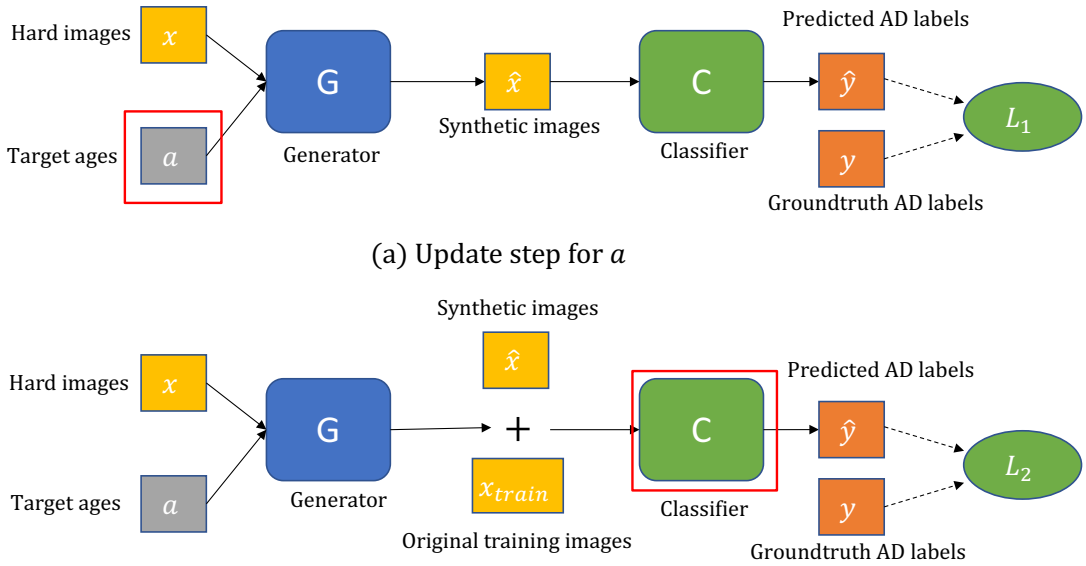
The use of Fourier encoding offers two advantages. First, in Chapter 5, we had to encode age and health state into two vectors and had to use two MLPs to embed the encoded vectors into the model. This may not be a big issue when the number of factors is small. However, if we want to extend the generative model to be conditioned on tens or hundreds of factors, the memory and computation costs will increase significantly. With Fourier encoding, we can encode all possible factors into a single vector, which offers more flexibility to scale the model to multiple conditional factors. Second, Fourier encoding allows us to compute the gradients with respect to the input vector $\mathbf{v}$ or certain elements of $\mathbf{v}$, since the encoding process is differentiable. As such, we replace the *ordinal encoding* with *Fourier encoding* in this chapter for all experiments. The generative model G takes $\mathbf{v}$ as input: $\hat{\mathbf{x}} = G(\mathbf{x}, \mathbf{v})$, where $\mathbf{v}$ represents target age and health state. However, as we only change the target age a in this chapter, we write the generative process as $\hat{\mathbf{x}} = G(\mathbf{x}, a)$ for simplicity.

6.3.3 Adversarial classification training with a pre-trained generator

Suppose we have a conditional generative model G and a classification model C. The aim is to make the best of G to improve the performance of C. To achieve this, we propose a procedure consisting of three steps. First, we pre-train the generative model G and the classification model C, respectively. Then we select a portion of training images for counterfactual synthesis. At last, we update the weights of C using the counterfactuals generated by G via an adversarial game. A schematic of the adversarial classification training is presented in

Encoding	SSIM	PSNR	MSE	PAD
Ord. Enc.	$0.79_{0.06}$	$26.1_{2.6}$	$0.042_{0.006}$	$4.2_{3.9}$
Four. Enc.	$0.79_{0.08}$	$25.9_{2.7}$	$0.043_{0.009}$	$4.1_{3.7}$

Table 6.1: Quantitative results of brain ageing model using *ordinal encoding* and *Fourier encoding*. For detail of the evaluation metrics please refer to text and Section 5.4.



(a) Update step for a

(b) Update step for C

Figure 6.1: A schematic of the adversarial classification training. We have a pre-trained generator G that takes a brain image $\mathbf{x}$ and a target age a as input, and outputs a synthetically aged image $\hat{\mathbf{x}}$ that corresponds to the target age a. We also have a classifier C that aims to predict the Alzheimer's Disease (AD) label for a given brain image. To utilise the generator G to improve the classifier C, we propose an adversarial training strategy that involves two steps: (a) the update step for the target age a, where we update a in the direction of maximising the classification loss. (Equation 6.5); (b) the update step for the classifier C, where we update C to minimise the classification error (Equation 6.7). Note here the weights of the generator are frozen, and we only update a and C alternatively.

Figure 6.1. Algorithm 1 summarises the steps of the method. Below we describe each step in detail.

Pre-training: We first pre-train the classification model C on the training dataset D_{train} : $\{\mathcal{X}_{train}, \mathcal{Y}_{train}\}$. Specifically, the generative model is trained using the same losses as in Chapter 5, except we use Fourier encoding to encode age and health state (diagnosis of AD). As a result, we obtain a pre-trained generative model G that can generate counterfactuals conditioned on given target ages a: $\hat{\mathbf{x}} = G(\mathbf{x}, a)$.

The classification model C is a deep neural network trained to predict the AD diagnosis from

brain images, optimised by minimising:

$$L_{pre-train} = \mathbb{E}_{\mathbf{x}\sim\mathcal{X}_{train},\mathbf{y}\sim\mathcal{Y}_{train}} L_s(C(\mathbf{x}), y), \tag{6.2}$$

where $L_s(.)$ is a supervised loss (cross-entropy loss in this chapter), $\mathbf{x}$ is a brain image, and y is its ground-truth AD label. In practice, we may have access to existing pre-trained models, and the pre-training step could be skipped.

Hard Sample Selection: Recent works [297, 298] suggest that training data samples have different *influence* on training a supervised model, *i.e.* some training data are harder for the task and could be more *effective* to train the model than others. In [297], the authors propose to up-sample, *i.e.* duplicate, the *hard* samples as a way to improve the model performance. In this chapter, we also assume that some data samples (and their counterfactuals) are more effective to improve the classier C, and we use a strategy similar to [297] to select these $hard$ samples. Specifically, we record the classification errors of all training samples for the pre-trained C at the end of training and then select $N = 100$ samples that have the highest errors. The selected *hard* samples are denoted as $D_{hard} : \{X_{hard}, Y_{hard}\}$.

Adversarial training: Previous works [268, 266, 299, 300, 301, 302] use generative models to randomly generate a number of synthetic data and then use these data to improve the downstream models. However, just as different training samples have different effects on training [297, 298], among all possible synthetic data, some could be more *useful* or *effective* to improve the downstream models than others. Similar to [297], we make an assumption that if a synthetic data sample is *hard* for the supervised model, then it tends to be more *effective* for training. We propose an adversarial game to find the *hard* synthetic data to improve C. A schematic of the adversarial training is presented in Figure 6.1.

Specifically, let us first define the classification loss for synthetic data as:

$$L_C = \mathbb{E}_{\mathbf{x}\sim X_{hard},y\sim Y_{hard}} L_s(C(\hat{\mathbf{x}}, y)), \tag{6.3}$$

where $\hat{\mathbf{x}}$ is a generated data sample that is conditioned on the target age a: $\hat{\mathbf{x}} = G(\mathbf{x}, a)$, and y is the ground-truth AD label for $\mathbf{x}$. Here we make an assumption that changing the target

age does not change the AD status, thus $\mathbf{x}$ and $\hat{\mathbf{x}}$ have the same AD label.

Since the encoding of age a is differentiable (see Section 6.3.2), we can obtain the gradients of L_C with respect to a as: $\nabla_a L_C = \nabla_a[L_s(C(G(\mathbf{x}, a)), y)]$, and update a in the direction of *maximising* L_C by:

$$\tilde{a} = a + \gamma_a \nabla_a L_C, \tag{6.4}$$

where γ_a is the step size (learning rate) for updating a, tuned to be 0.01 in this chapter. Formally, the optimization function of a can be written as:

$$L_1 = \max_a \mathbb{E}_{\mathbf{x}\sim X_{hard}, y\sim Y_{hard}} L_s(C(\hat{\mathbf{x}}), y), \tag{6.5}$$

Then we could obtain a set of synthetic data using the updated $\tilde{a}$, denoted as $D_{syn} : \{X_{syn}, Y_{syn}\}$.

The classifier C could be updated by minimising:

$$\min_C \mathbb{E}_{\mathbf{x}\sim X_{syn}, y\sim Y_{syn}} L_s(C(\mathbf{x}), y) = \mathbb{E}_{\mathbf{x}\sim X_{hard}, y\sim Y_{hard}} L_s(C(G(\mathbf{x}, \tilde{a})), y), \tag{6.6}$$

where $G(\mathbf{x}, \tilde{a})$ are the counterfactuals conditioned on the updated $\tilde{a}$. However, in practice we found that updating C only on the synthetic data could cause *catastrophic forgetting* [303], *i.e.* the classifier forgets what it learnt from the original training dataset (see Section 6.5.2 for more details). To prevent catastrophic forgetting, we adopt the same strategy as [297]: we update C on a combined dataset consisting of original training dataset and synthetic dataset: $\{X_{combined}, Y_{combined}\} = \{X_{train} \cup X_{syn}, Y_{train} \cup Y_{syn}\}$. Therefore, in practice, we update C by:

$$L_2 = \min_C \mathbb{E}_{\mathbf{x}\sim X_{combined}, y\sim Y_{combined}} L_s(C(\mathbf{x}), y). \tag{6.7}$$

The adversarial game is formulated by alternatively updating a and classifier C via Equation 6.5 and 6.7, respectively. In practice, to prevent a from going to unrealistic ages, we clip it to be in [60, 90] after every update.

Updating a vs. updating G: Note here the adversarial game is formulated between a and C, instead of G and C. One may wonder why not training G against C, *i.e.* updating G such

that it can produce *hard* counterfactuals to improve C. This may look like a good idea at first glance. However, by training G against C, we are allowing G to change its latent space without constraints to maintain image quality. In Section 6.5.4, we provide an ablation study that shows training G against C could result in unrealistic synthetic output, which can, in turn, hurt the performance of C.

6.4 Experimental setup

Data: For the experiments of this chapter, we use the *Alzheimer's Disease Neuroimaging Initiative* ADNI dataset, which is introduced in Section 2.5. We select 380 AD and 380 CN T1 volumes for our experiments, with 260 AD and 260 CN volumes as training data, 40 AD and 40 CN volumes as validation data, and 80 AD and 80 CN volumes as testing data. These volumes are from subjects between 60 and 90 yrs old.

Pre-processing: All volumetric data are skull-stripped using DeepBrain[2], and linearly regis-

[2]https://github.com/iitzco/deepbrain

Algorithm 1 Adversarial classification learning with a pre-trained generative model.

Input: Training set D_{train}; hyperparameter k, N; a pre-trained generator model G; the classifier model C.

Pre-training:

1. Train the classifier C on D_{train} for 100 epochs. (Equation 6.2)

Hard Sample Selection:

2. Select N samples from D_{train} that result in highest classification errors for C, denoted as D_{hard}.

Adversarial training:

3. Randomly initialise target ages a, and obtain initial synthetic data.
4. Update a in the direction to maximise classification error (Equation 6.5).
5. Obtain synthetic images with D_{hard} and the updated a, denoted as D_{syn}.
6. Update C to minimise the classification error on $D_{train} \cup D_{syn}$ (Equation 6.7).
7. Repeat 4,5,6 for k iterations.

tered to MNI 152 space using FSL-FLIRT [94]. We normalise brain volumes by clipping the intensities to $[0, V_{99.5}]$, where $V_{99.5}$ is the 99.5% largest intensity value within each volume, and then rescale the resulting intensities to the range $[-1, +1]$. We select the middle 60 axial slices from each volume and crop each slice to the size of $[208, 160]$. This results in 31200 training images, 4800 validation images and 9600 testing images.

Comparison methods: We compare with the following baselines:

- *Naïve*: We directly use the pre-trained classifier C for comparison.
- *Random Selection + Random Synthesis (RSRS)*: We randomly select N samples from the training data D_{train}, and then use the generator G to randomly generate $N_{synthesis} = 5$ synthetic samples per sample, denoted as D_{syn}. Then we train the classifier on the combined dataset $D_{train} \cup D_{syn}$ for $k = 5$ steps.
- *Hard Selection + Random Synthesis (HSRS)*: We select N hard samples from D_{train} based on their classification errors of C, and then use the generator G to randomly generate $N_{synthesis} = 5$ synthetic samples per hard sample, denoted as D_{syn}. Then we train the classifier on the combined dataset $D_{train} \cup D_{syn}$ for $k = 5$ steps.
- *Random Selection + Adversarial Training (RSAT)*: We randomly select N samples from D_{train}, and then use the adversarial training strategy to update the classifier C. The difference between RSAT and our approach is that we select hard samples for generating counterfactuals, while RSAT uses random samples.
- *Just Train Twice (JTT)*: We also compare with a recent work [297]. The idea of *JTT* is to record samples that are misclassified by the pre-trained classifier, obtaining an error set. Then they construct an upsampled dataset D_{up} that contain examples in the error set λ_{up} times and all other examples once. Finally, they train the classifier on the upsampled dataset D_{up}. In this chapter, we set $\lambda_{up} = 2$ as we found large λ_{up} results in bad performance. We also found the original learning rate (0.01) used for the second training stage results in a very bad performance and set it to be 0.00001.

Implementation details: The generator is trained the same way as in Chapter 5 (Section 5.4), except we replace *ordinal encoding* with *Fourier encoding*. We pre-train the classi-

fier for 100 epochs. The experiments are performed using Keras [223] and Tensorflow [304][3]. We train pre-trained classifiers C with Adam [224] with a learning rate of 0.00001 and decay of 0.0001. During adversarial learning, the step size of a is tuned to be 0.01, and the learning rate for C is 0.00001.

6.5 Results and Discussion

We start by presenting test accuracies of our approach and baselines, including the accuracies for different *age-diagnosis* test groups, demonstrating that our approach can improve the performance of the classifier C (Section 6.5.1). We then discuss our approach in a *continual learning* context and perform experiments to show that our approach can help alleviate *catastrophic forgetting* (Section 6.5.2). After that, we create a dataset with spurious correlations between *age* and *AD state* and show that our method might be able to break the spurious correlations (Section 6.5.3). Finally, we conclude with an ablation study where we train G against C (Section 6.5.4).

6.5.1 Main experiment

Here we first compare our procedure with baselines using the test accuracies of the classifiers. Note the weights of the pre-trained classifier C, and the pre-trained G generator are the same for all methods. The number of samples used to generate counterfactuals is set to $N = 100$ for our procedure and baselines *RSRS* and *HSRS*. For baselines *RSRS* and *HSRS*, we generate $N_{synthesis} = 5$ counterfactuals per sample. For a hard sample $x_i \in X_{hard}$, we randomly sample $a_i \sim U(\bar{a}_i, a_{max})$ as its initial target age, where $\bar{a}$ and a_{max} are the ground-truth age of x_i and the maximal age for all samples, respectively. The adversarial training of our approach and $RSAT$ is repeated for $k = 5$ epochs. Thus, for baselines *RSRS*, *HSRS*, *RSAT* and our approach, the number of synthetically augmented samples is 500. For JTT, there are 2184 samples misclassified by *C*. Table 6.2 presents the test accuracies of models trained with

[3]However, we do recommend using Pytorch [305] as it allows easier implementation to compute gradients w.r.t input.

Acc %	CN			AD			All
Age group	60-70yrs	70-80yrs	80-90yrs	60-70yrs	70-80yrs	80-90yrs	overall
Test group size	1540	1600	1660	1720	1540	1540	9600
Naïve	85.2	91.5	*70.7*	92.5	94.2	**97.1**	88.4
RSRS	86.0	90.4	*73.8*	87.3	95.1	90.0	87.0
HSRS	85.6	91.1	*80.4*	89.8	93.8	96.9	89.5
RSAT	86.1	93.1	*81.5*	91.8	96.0	95.7	90.6
JTT	83.9	**94.2**	*80.1*	**92.8**	90.8	93.7	89.2
Proposed	**86.4**	93.7	***83.4***	91.5	**96.5**	95.7	**91.1**

Table 6.2: Average test accuracies of models trained via our procedure and baselines. We first present the average test accuracies for different age groups with AD (column 2-4) or CN (column 5-7) and then present the average test accuracies for the whole testing set (column 8). For each method, the *worst-group* performance is shown in *italic*. For each age group, *i.e.* each column, the **best** performance is shown in **bold**. We also report the number of testing images for each age group.

our procedure and baseline. Specifically, we present the test accuracies for different test age groups with different AD diagnoses.

From Table 6.2 we can observe that our proposed procedure achieves the best overall test accuracy, followed by baseline *RSAT*. This demonstrates the advantage of adversarial training between the conditional factor (target age) a and the classifier. On top of that, it shows that selecting *hard* examples for creating augmented synthetic results helps, which is also demonstrated by the improvement of performance of *HSRS* over *Naïve*. We also observe that *JTT* [297] improves the classifier performance over *Naïve*, showing the benefit of up-sampling *hard* samples. In contrast, baseline *RSRS* achieves the lowest overall test accuracy, even lower than that of *Naïve*. This shows that randomly synthesising counterfactuals from randomly selected samples could result in synthetic images that are harmful to the classifier.

Furthermore, we observe that for all methods, the *worst-group* performances are achieved on the 80-90 CN group. A potential reason could be: as age increases, the brains shrink, and it is harder to tell if the ageing pattern is due to AD or caused by normal ageing. Nevertheless, we

AUC	60-70yrs	70-80yrs	80-90yrs	Overall
Naïve	0.954	0.968	0.903	0.931
RSRS	0.932	0.977	0.904	0.928
HSRS	0.958	0.975	**0.921**	0.954
RSAT	0.955	0.981	0.912	0.957
JTT	0.957	0.978	0.914	0.952
Proposed	**0.961**	**0.988**	0.917	**0.960**

Table 6.3: The test Area Under the ROC Curve (AUC) [16] values for all methods. We first present the AUC for different age groups (column 2-4), and then present the AUC for all testing data (column 5). For each group, the **best** results are shown in **bold**.

observe that for this *worst group*, our proposed method still achieves the best performance, followed by $RSAT$. This shows that adversarial training can be helpful to improve the performance of the classifier, especially for *hard* groups. The next best results are achieved by *HSRS* and *JTT*, which shows that finding hard samples and up-sampling or augmenting them was helpful to improve the *worst-group* performance. We also observe the improvement of *worst-group* performance for *RSRS* over *Naïve*, but the improvement is small compared to other baselines.

We also measure the Area Under Curve (AUC) values for all methods, as presented in Table 6.3. We can observe that our approach achieve the highest overall AUC results.

In summary, the quantitative results show that it is helpful to find and utilise *hard* counterfactuals for improving the classifier.

6.5.2 Adversarial classification training in a *continual learning* context

6.5.2.1 Connection to continual learning

Most previous works [268, 266, 299, 300, 301, 302] that used pre-trained deep generative models for augmentation focused on generating a large number of synthetic samples, and then merged the synthetic data with the original dataset and trained the downstream task

model (*e.g.* a classifier) on this augmented dataset. However, this requires training the task model from scratch, which could be inconvenient. For instance, if we suddenly decide to generate some new synthetic data for augmentation, we would have to retrain the task model from scratch. Furthermore, if the size of the original dataset is large, then the number of synthetic samples can be huge, which would make the training process extremely expensive and time-consuming. Thus, in practice, we need to consider cases where we aim to improve a pre-trained classifier with synthetic data but without retraining the whole model from scratch. We design the proposed procedure in such a way that allows us to use the pre-trained G to improve C flexibly.

In Section 6.5.1, after we obtain the synthetic set D_{syn}, we choose to update the classifier C on the augmented dataset $D_{syn} \cup D_{train}$, instead of D_{syn} (stage 7 in Algorithm 1). This is because re-training the classifier only on the D_{syn} would result in *catastrophic forgetting* [303], *i.e.* a phenomenon where deep neural networks tends to forget what it has learnt from previous data when being trained on new data samples. To alleviate catastrophic forgetting, efforts have been devoted to developing approaches to allow artificial neural networks to learn in a sequential manner [306, 307]. These approaches are known as *continual learning* [306, 308, 309], *lifelong learning* [310, 311], *sequential learning* [312, 313], or *incremental learning* [314, 315]. Despite different names and focuses, the main purpose of these approaches is to overcome catastrophic forgetting and to learn in a sequential manner.

If we consider the generated data as new samples, then the update of the pre-trained classifier C can be viewed as a *continual learning* problem, *i.e.* how to learn *new* knowledge from the synthetic set D_{syn} without forgetting *old* knowledge that is learnt from the original training data D_{train}. To alleviate catastrophic forgetting, we re-train the classifier on both the synthetic dataset D_{syn} and the original training dataset D_{train}. This strategy is known as *memory replay* in continual learning [316, 317] and was also used in other augmentation works [297]. The key idea is to store previous data in a *memory buffer* and *replay* the saved data to the model when training on new data. However, it could be expensive to store and revisit all the training data, especially when the data size is large [317]. In the next section, we perform experiments where we only provide a portion ($M\%$) of training data to the classifier when re-training with synthetic data (to simulate the *memory buffer*). We want to see whether *catastrophic*

Algorithm 2 Adversarial classification learning with D_{store}.

Input: Training dataset D_{train}; hyperparameter M, N, k; a pre-trained generator G; a pre-trained classifier model C.

Construct D_{store}:

1. Randomly select $M\%$ data from D_{train}, denoted as D_{store}.

Hard sample selection

2. Select N samples from D_{store} that result in highest classification errors for C, denoted as D_{hard}.

Adversarial training:

3. Randomly initialise target ages a, and obtain initial synthetic data.
4. Update a in the direction to maximise classification error (Equation 6.5).
5. Obtain synthetic images with D_{hard} and the updated a, denoted as D_{syn}.
6. Update C to minimise the classification error on $D_{store} \cup D_{syn}$ (Equation 6.7).
7. Repeat 4,5,6 for k iterations.

forgetting would happen or not when only a portion ($M\%$) of training data is provided, and if so, how much it affects the test accuracies.

6.5.2.2 Results when re-training with a portion ($M\%$) of training data

Suppose we have a pre-trained classifier C and a pre-trained generator G, and we want to improve C by using G for data augmentation. However, after pre-training, we only store $M\%$ ($M \in (0, 100]$) of the training dataset, denoted as D_{store}. During the adversarial training, we synthesise N samples using the generator G, denoted as D_{syn}. Then we update the classifier C on $D_{store} \cup D_{syn}$, using Equation 6.7 where $D_{combined} = D_{store} \cup D_{syn}$. The target ages are initialised and updated the same way as in Section 6.5.1. Algorithm 2 illustrates the procedure in this section.

Table 6.4 presents the test accuracies of our approach and baselines when M changes. For *Naïve-100*, the results are then same as in Table 6.2. For JTT, the original paper [297] re-trained the classifier using the whole training set. Here we first randomly select $M\%$ training samples as D_{store} and find misclassified data D_{mis} within D_{store} to up-sample, then we retrain

the classifier on the augmented set. We can observe that when M decreases, *catastrophic forgetting* happens for all approaches. However, our method suffers the least from catastrophic forgetting, especially when M is small. With $M = 20\%$ of training data for retraining, our approach achieves better results than *Naïve*. This might be because the adversarial training between a and C tries to detect what is missing in D_{store} and tries to recover the missing data by updating a towards those directions. We observe that *RSAT* achieves the second best results, only slightly worse than the proposed approach. Moreover, *HSRS* and *JTT* are more affected by catastrophic forgetting and achieve worse results. This might be because the importance of selecting *hard* samples declines as M decreases, since the D_{store} becomes smaller.

These results demonstrate that our approach could alleviate *catastrophic forgetting*. This could be helpful in cases where we want to utilise generative models to improve pre-trained classifiers (or other task models) without *revisiting* all the training data (a *continual learning* context).

6.5.2.3 Results when number of samples used for synthesis (N) changes

We also performed experiments where we changed N, *i.e.* the number of samples used for generating counterfactuals. Specifically, we set $M = 1$, *i.e.* only 1% of original training data

Acc %	$M\%$				
Methods	1	10	20	50	100
Naïve	N/A	N/A	N/A	N/A	88.4
HSRS	75.6	81.4	84.5	87.4	89.5
RSAT	84.2	85.8	87.2	88.6	90.6
JTT	77.3	82.3	85.1	88.1	89.2
Proposed	84.8	86.8	88.5	89.4	91.1

Table 6.4: Test accuracies of our approach and baselines when the ratio of the size D_{store} vs. the size of D_{train} changes. We can observe the decreases of test accuracies when M decreases, which was due to the effect of *catastrophic forgetting*.

are used for re-training C, to see how many synthetic samples are needed to maintain good accuracy, especially when there are only a few training data stored in D_{store}. This is to see how *efficient* the synthetic samples are in terms of training C and alleviating *catastrophic forgetting*. The results are presented in Table 6.5.

From Table 6.5, we can observe that the best results are achieved by our method, followed by *RSAT*. Even with only one sample for synthesis, our method could still achieve a test accuracy of 80%. This is probably because the adversarial training of a *vs.* C guides G to generate *hard* counterfactuals, which are efficient to train the classifier. The results demonstrate that our approach could help alleviate *catastrophic forgetting* even with a small number of synthetic samples used for augmentation. This experiment could also be viewed as a measurement of the *sample efficiency*, *i.e.* how efficient a synthetic sample is in terms of re-training a classifier.

6.5.3 Can the proposed procedure alleviate *spurious correlations*?

Spurious correlation occurs when two factors appear to be correlated to each other but in fact they are not [274]. Spurious correlation could affect the performance of deep neural networks and has been actively studied in computer vision field [318, 297, 318, 319, 320, 287] and in medical imaging analysis field [321, 322]. For instance, suppose we have an dataset of *bird* and *bat* photos. For *bird* photos, most backgrounds are *sky*. For *bat* photos, most backgrounds are *cave*. If a classifier learns this spurious correlation, *e.g.* it classifies a photo as *bird* as long as the background is *sky*, then it will perform poorly on images where *bats* are flying in the *sky*. In this section, we investigate if our approach could correct such *spurious correlations* by changing a to generate hard counterfactuals.

acc %	*N*			
Methods	1	10	50	100
HSRS	65.4	71.0	73.4	75.6
RSAT	81.3	82.1	83.2	84.2
Proposed	82.1	82.9	84.1	84.6

Table 6.5: Test accuracies when N changes ($M = 1$) of our approach and baselines.

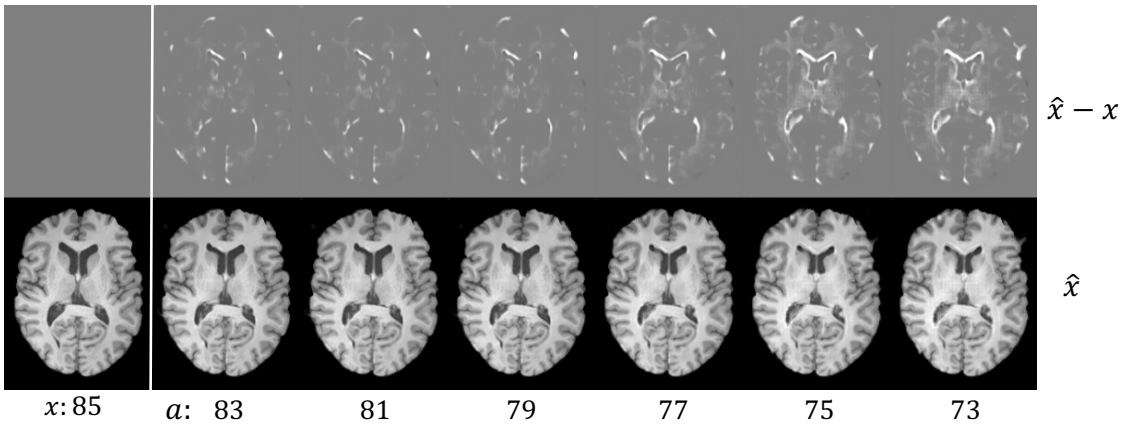


Figure 6.2: Example results of brain *rejuvenation* for an image (x) of a 85 year old CN subject. We synthesise *rejuvenated* images $\hat{x}$ at different target ages a. We also show the differences between $\hat{x}$ and x, $\hat{x} - x$. For more details see text.

Specifically, we create a dataset where 7860 images between 60 and 75 yrs old are AD, and 7680 images between 75 and 90 yrs old are healthy, denoted as $D_{spurious}$. This is to construct *spurious correlations*: $young \rightarrow AD$ and $old \rightarrow CN$ (in reality older people have higher chances of getting AD). Then we pre-train C on $D_{spurious}$, and select 50 AD and 50 CN subjects for generating synthetic data. Since the brain ageing model proposed in Chapter 5 only considered simulating *ageing* process, but did not consider *rejuvenation* where brains are transformed from old to young. To be able to use *old CN* data, we pre-train another generator model in the opposite direction, *i.e.* generating *younger* brain images from old ones. As a result, we obtain two pre-trained generators, denoted as G_{ageing} and $G_{rejuvenation}$, where G_{ageing} is the generator used in previous sections and $G_{rejuvenation}$ is trained to simulate the *rejuvenation* process. Figure 6.2 illustrates example visual results of $G_{rejuvenation}$, where we synthesise the *rejuvenated* images $\hat{x}$ for a 85-year-old CN subject x at different target ages a. From Figure 6.2 we observe that although the brain ageing model in Chapter 5 was designed for brain *ageing*, it can be used to simulate brain rejuvenation with high-quality results.

After we obtain G_{ageing} and $G_{rejuvenation}$, we select 50 CN and 50 AD images from $D_{spurious}$ that result in highest training errors, denoted as D_{hard}. Note the selected CN images are between 75 and 90 yrs old, and the AD images are between 60 and 75 yrs old. Then we generate synthetic images from D_{hard} using $G_{rejuvenation}$ for old CN samples and G_{ageing} for

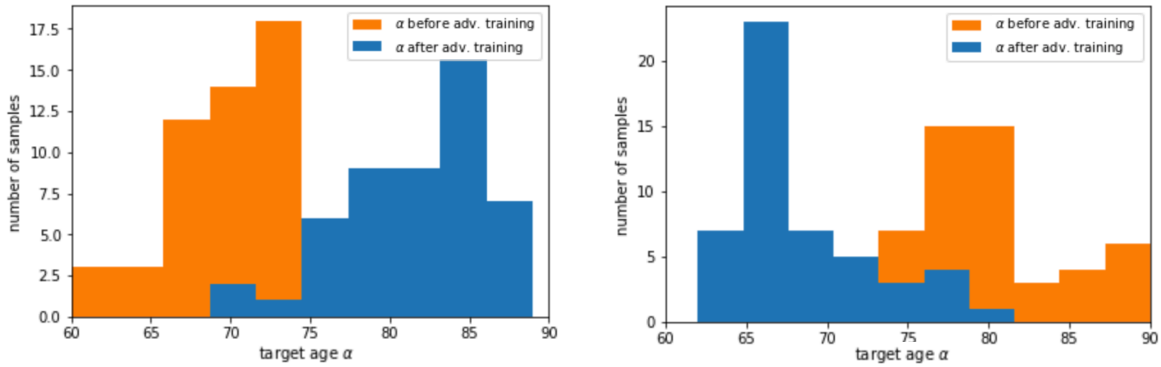
Acc %	CN		AD		
Methods	60-75yrs	75-90yrs	60-75yrs	75-90yrs	Overall
Naïve	*40.9*	81.6	**95.1**	45.7	67.0
HSRS	*60.7*	85.3	81.1	67.2	75.0
JTT	50.5	**88.4**	85.5	*40.7*	67.9
proposed	***73.1***	83.4	81.5	**75.8**	**79.0**

Table 6.6: Test accuracies for our procedure and baselines when C pre-trained on $D_{spurious}$. We first present the average test accuracies for different age groups with CN diagnosis (column 2-3) or AD (column 4-5), and then present the average test accuracies for the whole testing set (column 6). For each method, the *worst-group* performance is shown in *italic*. For each age group, *i.e.* each column, the **best** performance was shown in **bold**. For more details see text.

young AD samples. The target ages a are initialized as their ground-truth ages. Finally, we perform the adversarial training between a and the classifier C. Here we want to see if the adversarial training can detect the *spurious correlations* purposely created by us, and more importantly, we want to see if the adversarial training between a and C can *break* the spurious correlations.

Table 6.6 presents the test accuracies of our approach and baselines. For *Naïve*, we directly use the classifier C pre-trained on $D_{spurious}$. For *HSRS*, we randomly generate synthetic samples from D_{hard} for augmentation. For *JTT*, we simply select mis-classified samples from $D_{spurious}$ and up-sample these samples.

We can observe from Table 6.6 that the pre-trained C on $D_{spurious}$ (*Naïve*) achieves much worse performance (67% accuracy) compared to that of Table 6.2 (88.4% accuracy). Specifically, it tends to misclassify *young CN* images as AD and misclassify *old AD* images as CN. This is likely due to the spurious correlations that we purposely create in $D_{spurious}$: $young \rightarrow AD$ and $old \rightarrow CN$. We notice that for *Naïve*, the test accuracies of AD groups are higher than that of CN groups. This is likely due to the fact we have more AD training data, and the classifier is biased to classify a subject to AD. This can be viewed as another *spurious correlation*. Overall, we observe that our method achieves the best results, followed



(a) Target ages a for AD subjects in D_{hard}.

(b) Target ages a for CN subjects in D_{hard}.

Figure 6.3: Histograms of target ages a before and after adversarial training: (a) the histogram of a for the 50 AD subjects in D_{hard}; (b) the histogram of a for the 50 CN subjects in D_{hard}. Here we show histograms of a before (in orange) and after (in blue) the adversarial training.

by *HSRS*. This shows that the synthetic results generated by the generators are helpful to alleviate the effect of *spurious correlations* and improve downstream tasks. The improvement of our approach over *HSRS* is due to the adversarial training between a and C, which guides the generator to produce hard counterfactuals. We observe *JTT* does not improve the test accuracies significantly. A potential reason is that *JTT* tries to find 'hard' samples in the training dataset. However, in this experiment, the 'hard' samples should be *young CN* and *old AD* samples which do not exist in the training dataset $D_{spurious}$. By contrast, our procedure could guide G to generate these samples, and *HSRS* could create these samples by random chance.

Figure 6.3 plots the histograms of the target ages a before and after the adversarial training. From Figure 6.3 we can observe that the adversarial training pushes a towards the *hard* direction, which could alleviate the spurious correlations. For instance, in $D_{spurious}$ and D_{hard} the AD subjects are all in the *young* group, *i.e.* 60-75 yrs old, and the classifier learns the *spurious correlation*: $young \rightarrow AD$, but in Figure 6.3 (a) we can observe that the adversarial training learns to generate AD synthetic images in the range of 75-90 yrs old. These *old AD* synthetic images can help alleviate the spurious correlation and improve the performance of C. Similarly, we can observe a are pushed towards *young* for CN subjects in Figure 6.3 (b).

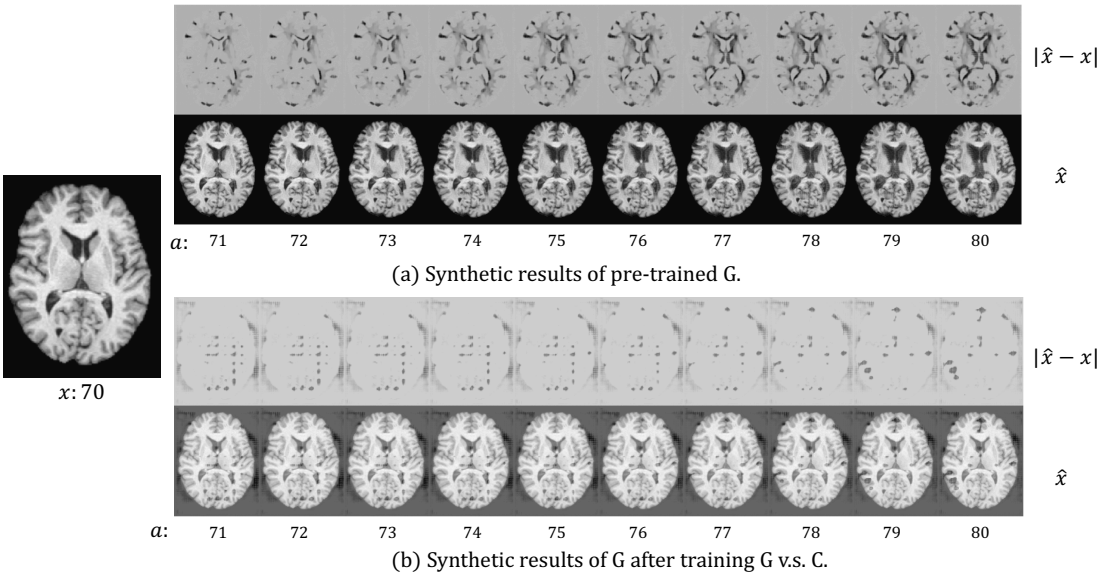


Figure 6.4: The synthetic results for a healthy (CN) subject x at age 70: (a) the results of the pre-trained G, *i.e.* before we train G against C; (b) the results of G after we train G against C. We synthesise aged images $\hat{x}$ at different target ages a. We also visualise the difference between x and $\hat{x}$, $|\hat{x} - x|$. For more details see text.

6.5.4 Ablation study: train *G* against *C*

We choose to formulate an adversarial game between the conditional generative factor a (the target age) and the classifier C, instead of between the generator G and the classifier C. This is because we are concerned that an adversarial game between G and C could result in unrealistic outputs of G. In this section, we perform an experiment to investigate this.

Specifically, we define an optimization function:

$$L_G = \max_G \mathbb{E}_{\mathbf{x} \sim X_{train}, \mathbf{y} \sim Y_{train}} L_s(C(G(\mathbf{x}, a)), y), \tag{6.8}$$

where we aim to train G in the direction of maximising the loss of the classifier C on the synthetic data $G(\mathbf{x}, a)$.

After every update of G, we construct a synthetic set D_{syn} by generating 100 synthetic images from D_{train}, and update C on $D_{train} \cup D_{syn}$ via Equation 6.7. The adversarial game G vs. C is formulated by alternatively optimising Equation 6.8 and 6.7 for 10 epochs.

In Figure 6.4, we present the synthetic brain ageing progression of a CN subject before and after the adversarial training of G vs. C. We can observe that after the adversarial training, the generator G produces unrealistic results. This could be because there is no loss or constraint to prevent the generator G from producing low-quality results. The adversarial game only requires the generator G to produce images that are hard for the classifier C, and naturally, images of low quality would be hard for C. A potential solution could be to involve a GAN loss with a discriminator to improve the output quality, but this would make the training much more complex and require more memory and computations. We also measure the test accuracy of the classifier C after training G against C to be 81.6%, which is much lower than the *Naïve* method (88.4%) and our approach (91.1%) in Table 6.2. The potential reason is that C is misled by the unrealistic samples generated by G.

6.6 Summary

In this chapter, we presented a simple procedure to utilise conditional generative models for downstream tasks. The proposed procedure requires a pre-trained conditional generative model and builds an adversarial game between the downstream task model and the generative conditional factor. To demonstrate this strategy, we choose the brain ageing synthesis model in Chapter 5 as the generative model and focus on the problem of classifying Alzheimer's Disease. We presented quantitative results showing that our procedure can improve the performance of the AD classifier. We also discussed the proposed procedure from a *continual learning* perspective and presented results showing that our procedure could help alleviate *catastrophic forgetting* when re-training a model with synthetic data. Besides, we constructed a dataset where we purposely created *spurious correlations* and showed that our method could alleviate the effect of spurious correlations.

Chapter 7
Conclusion and Future Directions

The final chapter concludes this thesis by summarising thesis contributions and discussing limitations and future directions.

7.1 Summary

This thesis focuses on deep learning methods for medical image synthesis. We propose new approaches for pseudo healthy synthesis and brain ageing synthesis. We also define quantitative metrics to measure the quality of synthetic images when there are no ground-truth images available. Finally, we propose an adversarial training strategy to utilise pre-trained generative models for downstream tasks, which has the potential to be applied to existing pre-trained generative models.

Chapter 4 focuses on *pseudo healthy synthesis*, i.e. the creation of subject-specific 'healthy' images from pathological ones. We propose a model that can generate realistic 'healthy' images with maintaining subject identity. The proposed method can be trained in *paired* and *unpaired setting*. To measure the quality of the synthetic images, we propose several metrics focusing on *healthiness*, *identity preservation* and *deformation correction*. We also include a human study to evaluate the results. Both qualitative and quantitative results show that the proposed method can produce realistic 'healthy' counterfactuals while preserving subject identity.

Chapter 5 focuses on *brain ageing synthesis* where the goal is to simulate how a brain would look like when age increases. We propose a deep learning model that can learn the brain ageing progression without longitudinal data. We conduct a series of experiments to evaluate the quality of synthetically aged images, including longitudinal evaluation and age accuracy measurement by a pre-trained age predictor. The qualitative and quantitative results show

that our approach can generate realistic aged brain images that preserve subject identity and are conditioned on given target age and health status.

Chapter 6 aims to find ways to utilise a pre-trained generative model for improvements on downstream tasks. Specifically, we focus on the classification of Alzheimer's Disease and utilise the brain ageing model in Chapter 5 to improve the classification performance. We also consider the use of generative models in a *continual learning* context and show that the proposed procedure could alleviate *catastrophic forgetting*. Furthermore, we demonstrate that the proposed approach could be used to correct *spurious correlations*. The results show that the proposed strategy can improve the classification performance and has the potential to be applied to other generative models for different tasks.

7.2 Limitations and Future Directions

This thesis has some limitations that can inspire future directions. Due to the GPU memory limit, our models remain 2D. A general future direction could be to extend these models for 3D medical images as 3D medical images contain more information. Moreover, our models focus on single modality medical data, while in practice, different modalities contain information that is complementary to each other. Thus, it is beneficial to build multi-input multi-output models that can work on multi-modality data, but this will further increase parameter space. Furthermore, the proposed generative models only consider one specific pathology. It could be beneficial to develop models that take multiple pathology and clinical factors into account. Ablation studies of the effect of loss weights could be provided. Although in this thesis we focus on brain MRI images, the proposed methods can be applied to other organs. Below we discuss limitations and future directions specific for each chapter.

Chapter 4: We see several avenues for future works. Metrics that enforce or even measure identity is a topic of considerable interest in computer vision [200]. One of our proposed metrics aimed to assess whether the subject identity has been preserved in synthetic 'healthy' images, while another metrics assessed if deformation caused by disease was recovered. Analysis combining these two metrics could assess the preservation of identity even when deformation was corrected, which is suited for cases where disease globally affects an

image. Further lines of improvement involve better methods to measure the null hypothesis (e.g. perhaps by artificially creating images from the healthy class that seem to be distorted). In addition, we do see that human evaluation is useful, although challenging as it requires expertise. Moreover, most clinical neurologists do not evaluate medical images in isolation but rather consider them in combination with other medical information in order to make a diagnostic decision. Nevertheless, we have performed a human experiment involving a neurologist, which best adhered to a blinded workflow. However, better evaluation schemes could be proposed, which is seen as a future direction. We also see a future opportunity in creating a large benchmark study that amasses expert evaluations, which are used to learn combinations of several quantitative, yet easy to obtain, numerical metrics that can act as surrogates to human evaluations.

Chapter 5: The proposed method has several potential applications. For example, a common problem in longitudinal studies is missing data due to patient dropout or poor-quality scans. The proposed method offers an opportunity to impute missing data at any time point. Furthermore, when there is insufficient longitudinal training data, the proposed method can be used to include cross-sectional data within a study. The simple experiment in Section 5.5.2.3 shows a glimpse of this potential. This, in turn, will make a further clinical analysis of ageing patterns, e.g. to evaluate the incidence of white matter hyperintensities [323], and large studies into neurodegenerative diseases, possible. Finally, from an AI perspective, we advocated earlier in this chapter about the importance of capturing and understanding the current state from a machine learning perspective. In fact, recently, this has been cast in causal inference and counterfactual setting [196]. While our work did not explicitly use a causal inference framework, our generated outputs can be seen as counterfactuals. This is evidenced by the experiments we performed in Chapter 6.

The notion of subject identity is context-specific, and we do note that others in the literature also follow the same simple assumptions we make. We do agree, though, that identity should be defined as what remains invariant under ageing and neurodegenerative disease. Although we used several losses to help preserve the subject identity of synthetic aged images, there is no guarantee that a subject's identity will be preserved, and new losses or mechanisms that could further improve identity preservation will be of high value. Unfortunately, without

access to large data where we exhaustively explore all possible combinations of variables that we want to be equivalent (to identity) or invariant (to age, pathology), preservation of identity can only be proxied. Although the proposed model only considers predicting older brain images from young ones, we show in Chapter 6 that it is possible to perform brain rejuvenation. The proposed method allows for a change of health status between input and output images. However, it does not model change of health state in between input and output. This is a common limitation of current works in this area [191, 196, 188]. A potential solution is recursive image synthesis: generating a suitable intermediate image before generating the desired target output of older age and state. Advances in architectures that improve image quality will enable such recursive image generation in the future. Conditioning mechanisms that reliably embed prior information into neural networks enabling finer control over the outputs of models, are of considerable interest in deep learning. In this chapter we design a simple yet effective way to encode both *age* (continuous) and *AD status* (ordinal) factors into the image generation network. However, as classification of MCI is challenging, the use of further (fine-grained) clinical information (e.g. clinical score) to reflect health status can be of benefit. Incorporating additional clinical variables, e.g. gender, genotypes, etc., can become inefficient with our current approach as it may involve more dense layers. While new techniques are available [324, 325, 326, 327] and some prior examples on few conditioning variables [328] or disentanglement [216] are promising, their utility in integrating clinical variables and replacing the need for ordinal pre-encoding of continuous or ordinal variables with imaging data is under investigation. In Chapter 6, we attempt *Fourier encoding* which can encode this information in a flexible way. Furthermore, here we focus on the use of cross-sectional data to train a model to predict aged brain images. If longitudinal data are also available, e.g. within a large study aggregating several data sources, model performance could be further improved by introducing supervised losses; however, adding more losses requires that they are well balanced –a known problem in semi-supervised learning [329].

Chapter 6: The proposed procedure has the potential to be applied to other pre-trained conditional generators on other datasets, *e.g.* face ageing synthesis [330, 201], face attributes editing [331]. It would be helpful to show results with more methods and more datasets. We must admit that the comparison in Section 6.5.3 is not strictly 'fair'. The generators G_{ageing} and $G_{rejuvenation}$ are pre-trained on D_{train} where the spurious correlations did not exist. How-

ever, the purpose is to show that a well-trained generative model can be used to alleviate *spurious correlations*. In practice, we could just use some existing public pre-trained generators for downstream tasks, *e.g.* StyleGAN[332, 333][1]. There were some recent works that tried to develop generative models that were less affected by spurious correlations [286, 334]. Thus even training on the same *spurious correlated* datasets, generative models might still be able to improve downstream tasks by alleviating the effect of spurious correlations. Nevertheless, we leave this as an avenue for future improvements by the community and us. The way we updated the conditional factor (target age) with gradients was preliminary and could be improved. Instead of directly adding gradients to a (Equation 6.5), we could use some optimization algorithms, *e.g.* Adam [224], to update the target age a. Instead of classifying between AD and CN subjects, we can try a more complex task, *i.e.* classifying AD, MCI and CN subjects, or classifying between stable MCI (sMCI) and progressive MCI (pMCI) subjects. Furthermore, instead of a VGG-based classifier, we can try to use some state-of-the-art AD classification models [335]. In this chapter, we consider a continuous scalar as the conditional factor, but we can also try to use adversarial training when the conditional factor is a discrete value or an image. Moreover, we can try to combine conventional data augmentation techniques with our approach.

Outlook: There are several general future directions in medical image analysis worth investigating. First, the utility of synthetic images should be taken into account. Currently, most medical synthetic works use the generated medical images as data augmentations to train downstream tasks. However, this may not fully exploit the value of these generated images. In some cases, these generated images are, in nature, counterfactuals. These counterfactuals can be helpful for causal inference. Second, the reliability and explainability of the synthesis models (and other task models) need to be explored. To make clinicians trust AI, AI systems' diagnosis results must be explainable. Otherwise, it is hard to convince clinicians and perhaps patients to trust the results coming from a black box, especially when the result relates to people's health or even life. Finally, an AI system that can diagnose various diseases is always valuable. At this moment, researchers in this field only focus on one specific disease in one project due to resource and time limits. However, in practice, clinicians need to consider

[1]Unfortunately, our GPU memory does not support the use of pre-trained StyleGANs available at `https://github.com/NVlabs/stylegan`.

the possibility of various diseases when diagnosing. With the increase in computation power and more availability of medical data, a general system that can provide more utility should be considered.

7.3 Epilogue

This thesis proposed two conditional generative models that are conditioned on discrete and continuous factors, respectively. Besides, a strategy to utilise these generative models for downstream tasks is proposed. We believe the generative models and the way to utilise them can be useful for amany research fields where data are insufficient. Also, although we did not explicitly focus on causal counterfactuals, our generative models are, in nature, counterfactual models. Thus, these models could be useful for studies in causality and counterfactual. We believe the work in this thesis can prove helpful for further research in the field of medical image analysis.

References

[1] C. B. Paschal and H. D. Morris, "K-space in the clinic," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 19, no. 2, pp. 145–159, 2004.

[2] J. R. Taylor, N. Williams, R. Cusack, T. Auer, M. A. Shafto, M. Dixon, L. K. Tyler, and R. N. Henson, "The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample," *NeuroImage*, vol. 144, pp. 262–9, 2017.

[3] H. B. Van der Worp and J. van Gijn, "Acute ischemic stroke," *New England Journal of Medicine*, vol. 357, no. 6, pp. 572–579, 2007.

[4] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[5] R. C. Petersen, P. Aisen, L. A. Beckett, M. Donohue, A. Gamst, D. J. Harvey, C. Jack, W. Jagust, L. Shaw, A. Toga, *et al.*, "Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, IEEE, 2017.

[7] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *ICLR*, 2017.

[8] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville, "Adversarially learned inference," *ICLR*, 2017.

[9] D. H. Ye, D. Zikic, B. Glocker, A. Criminisi, and E. Konukoglu, "Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 606–613, Springer, 2013.

[10] M. I. Miller, G. E. Christensen, Y. Amit, and U. Grenander, "Mathematical textbook of deformable neuroanatomies," *Proceedings of the National Academy of Sciences*, vol. 90, no. 24, pp. 11944–11948, 1993.

[11] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni, "Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study," *Medical Image Analysis*, p. 101952, 2021.

[12] C. F. Baumgartner, L. M. Koch, K. Can Tezcan, J. Xi Ang, and E. Konukoglu, "Visual feature attribution using wasserstein GANs," in *CVPR*, pp. 8309–19, 2018.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[15] Xia, Tian and Chartsias, Agisilaos and Tsaftaris, Sotirios A and Alzheimer's Disease Neuroimaging Initiative and others, "Consistent brain ageing synthesis," in *MICCAI*, pp. 750–758, Springer, 2019.

[16] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[19] A. F. Frangi, S. A. Tsaftaris, and J. L. Prince, "Simulation and synthesis in medical imaging," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 673–679, 2018.

[20] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, *et al.*, "ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI," *Medical image analysis*, vol. 35, pp. 250–269, 2017.

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

[22] F. Bloch, "Nuclear induction," *Physical review*, vol. 70, no. 7-8, p. 460, 1946.

[23] E. M. Purcell, H. C. Torrey, and R. V. Pound, "Resonance absorption by nuclear magnetic moments in a solid," *Physical review*, vol. 69, no. 1-2, p. 37, 1946.

[24] P. C. Lauterbur, "Image formation by induced local interactions: examples employing nuclear magnetic resonance," *nature*, vol. 242, no. 5394, pp. 190–191, 1973.

[25] P. Mansfield, "Multi-planar image formation using nmr spin echoes," *Journal of Physics C: Solid State Physics*, vol. 10, no. 3, p. L55, 1977.

[26] R. Hawkes, G. Holland, W. Moore, and B. Worthington, "Nuclear magnetic resonance (nmr) tomography of the brain: a preliminary clinical assessment with demonstration of pathology," *Journal of Computer Assisted Tomography*, vol. 4, no. 5, pp. 577–586, 1980.

[27] F. Smith, J. Hutchison, J. Mallard, G. Johnson, T. W. Redpath, R. Selbie, A. Reid, and C. Smith, "Oesophageal carcinoma demonstrated by whole-body nuclear magnetic resonance imaging.," *Br Med J (Clin Res Ed)*, vol. 282, no. 6263, pp. 510–512, 1981.

[28] B. U. Forstmann, M. C. Keuken, and A. Alkemade, "An introduction to human brain anatomy," in *An introduction to model-based cognitive neuroscience*, pp. 71–89, Springer, 2015.

[29] J. D. Schmahmann, J. Doyon, M. Petrides, A. C. Evans, and A. W. Toga, *MRI atlas of the human cerebellum*. Academic press, 2000.

[30] G. Paxinos and X.-F. Huang, *Atlas of the human brainstem*. Elsevier, 2013.

[31] A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison, and C. J. Murray, "Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data," *The lancet*, vol. 367, no. 9524, pp. 1747–1757, 2006.

[32] L. B. Goldstein, R. Adams, M. J. Alberts, L. J. Appel, L. M. Brass, C. D. Bushnell, A. Culebras, T. J. DeGraba, P. B. Gorelick, J. R. Guyton, *et al.*, "Primary prevention of ischemic stroke: A guideline from the american heart association/american stroke association stroke council: Cosponsored by the atherosclerotic peripheral vascular disease interdisciplinary working group; cardiovascular nursing council; clinical cardiology council; nutrition, physical activity, and metabolism council; and the quality of care and outcomes research interdisciplinary working group: The american academy of neurology affirms the value of this guideline.," *Stroke*, vol. 37, no. 6, pp. 1583–1633, 2006.

[33] P. Rothwell, A. Coull, L. Silver, J. Fairhead, M. Giles, C. Lovelock, J. Redgrave, L. Bull, S. Welch, F. Cuthbertson, *et al.*, "Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (oxford vascular study)," *The Lancet*, vol. 366, no. 9499, pp. 1773–1783, 2005.

[34] V. L. Feigin, C. M. Lawes, D. A. Bennett, and C. S. Anderson, "Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century," *The lancet neurology*, vol. 2, no. 1, pp. 43–53, 2003.

[35] J. Rojas, M. Zurru, M. Romano, L. Patrucco, and E. Cristiano, "Acute ischemic stroke and transient ischemic attack in the very old–risk factor profile and stroke subtype between patients older than 80 years and patients aged less than 80 years," *European Journal of Neurology*, vol. 14, no. 8, pp. 895–899, 2007.

[36] J. L. Fisher, J. A. Schwartzbaum, M. Wrensch, and J. L. Wiemels, "Epidemiology of brain tumors," *Neurologic clinics*, vol. 25, no. 4, pp. 867–890, 2007.

[37] M. Wrensch, Y. Minn, T. Chew, M. Bondy, and M. S. Berger, "Epidemiology of primary brain tumors: current concepts and review of the literature," *Neuro-oncology*, vol. 4, no. 4, pp. 278–299, 2002.

[38] T. Vos, A. D. Flaxman, M. Naghavi, R. Lozano, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, V. Aboyans, *et al.*, "Years lived with disability (ylds) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010," *The lancet*, vol. 380, no. 9859, pp. 2163–2196, 2012.

[39] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, "The hallmarks of aging," *Cell*, vol. 153, no. 6, pp. 1194–1217, 2013.

[40] J. L. Horn and R. B. Cattell, "Refinement and test of the theory of fluid and crystallized general intelligences.," *Journal of educational psychology*, vol. 57, no. 5, p. 253, 1966.

[41] M. Ziegler, E. Danay, M. Heene, J. Asendorpf, and M. Bühner, "Openness, fluid intelligence, and crystallized intelligence: Toward an integrative model," *Journal of Research in Personality*, vol. 46, no. 2, pp. 173–183, 2012.

[42] J. L. Horn and R. B. Cattell, "Age differences in fluid and crystallized intelligence," *Acta psychologica*, vol. 26, pp. 107–129, 1967.

[43] P. Ghisletta, P. Rabbitt, M. Lunn, and U. Lindenberger, "Two thirds of the age-based changes in fluid and crystallized intelligence, perceptual speed, and memory in adulthood are shared," *Intelligence*, vol. 40, no. 3, pp. 260–268, 2012.

[44] I. J. Deary, J. Corley, A. J. Gow, S. E. Harris, L. M. Houlihan, R. E. Marioni, L. Penke, S. B. Rafnsson, and J. M. Starr, "Age-associated cognitive decline," *British medical bulletin*, vol. 92, no. 1, pp. 135–152, 2009.

[45] R. A. Kievit, S. W. Davis, J. Griffiths, M. M. Correia, R. N. Henson, *et al.*, "A watershed model of individual differences in fluid intelligence," *Neuropsychologia*, vol. 91, pp. 186–198, 2016.

[46] A. K. Barbey, "Network neuroscience theory of human intelligence," *Trends in cognitive sciences*, vol. 22, no. 1, pp. 8–20, 2018.

[47] A. M. Fjell and K. B. Walhovd, "Structural brain changes in aging: courses, causes and cognitive consequences," *Reviews in the Neurosciences*, vol. 21, no. 3, pp. 187–222, 2010.

[48] R. Peters, "Ageing and the brain," *Postgraduate Medical Journal*, vol. 82, no. 964, pp. 84–88, 2006.

[49] A. J. Parkin, *Memory and amnesia: An introduction*. Psychology Press, 2013.

[50] L. R. Squire and S. M. Zola, "Episodic memory, semantic memory, and amnesia," *Hippocampus*, vol. 8, no. 3, pp. 205–211, 1998.

[51] R. Cabeza, "Neuroscience frontiers of cognitive aging: Approaches to cognitive neuroscience of aging," *New frontiers in cognitive aging*, pp. 179–196, 2004.

[52] C. Lustig and R. L. Buckner, "Preserved neural correlates of priming in old age and dementia," *Neuron*, vol. 42, no. 5, pp. 865–875, 2004.

[53] D. J. Madden, I. J. Bennett, and A. W. Song, "Cerebral white matter integrity and cognitive aging: contributions from diffusion tensor imaging," *Neuropsychology review*, vol. 19, no. 4, p. 415, 2009.

[54] D. Tomasi and N. D. Volkow, "Aging and functional brain networks," *Molecular psychiatry*, vol. 17, no. 5, pp. 549–558, 2012.

[55] M. M. Esiri, "Ageing and the brain," *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, vol. 211, no. 2, pp. 181–187, 2007.

[56] C. K. Tamnes, K. B. Walhovd, A. M. Dale, Y. Østby, H. Grydeland, G. Richardson, L. T. Westlye, J. C. Roddey, D. J. Hagler Jr, P. Due-Tønnessen, *et al.*, "Brain development and aging: overlapping and unique patterns of change," *Neuroimage*, vol. 68, pp. 63–74, 2013.

[57] E. V. Sullivan, L. Marsh, D. H. Mathalon, K. O. Lim, and A. Pfefferbaum, "Age-related decline in mri volumes of temporal lobe gray matter but not hippocampus," *Neurobiology of aging*, vol. 16, no. 4, pp. 591–606, 1995.

[58] Y. Taki, R. Goto, A. Evans, A. Zijdenbos, P. Neelin, J. Lerch, K. Sato, S. Ono, S. Kinomura, M. Nakagawa, *et al.*, "Voxel-based morphometry of human brain with age and cerebrovascular risk factors," *Neurobiology of aging*, vol. 25, no. 4, pp. 455–463, 2004.

[59] C. Van Petten, E. Plante, P. S. Davidson, T. Y. Kuo, L. Bajuscak, and E. L. Glisky, "Memory and executive function in older adults: relationships with temporal and prefrontal gray matter volumes and white matter hyperintensities," *Neuropsychologia*, vol. 42, no. 10, pp. 1313–1335, 2004.

[60] G. Kalpouzos, G. Chételat, J.-C. Baron, B. Landeau, K. Mevel, C. Godeau, L. Barré, J.-M. Constans, F. Viader, F. Eustache, *et al.*, "Voxel-based mapping of brain gray matter volume and glucose metabolism profiles in normal aging," *Neurobiology of aging*, vol. 30, no. 1, pp. 112–124, 2009.

[61] K. M. Kennedy and N. Raz, "Aging white matter and cognition: differential effects of regional variations in diffusion properties on memory, executive functions, and speed," *Neuropsychologia*, vol. 47, no. 3, pp. 916–927, 2009.

[62] R. A. Charlton, T. Barrick, D. McIntyre, Y. Shen, M. O'sullivan, F. Howe, C. Clark, R. Morris, and H. Markus, "White matter damage on diffusion tensor imaging correlates with age-related cognitive decline," *Neurology*, vol. 66, no. 2, pp. 217–222, 2006.

[63] R. A. Charlton, S. Landau, F. Schiavone, T. Barrick, C. Clark, H. Markus, and R. Morris, "A structural equation modeling investigation of age-related variance in executive function and dti measured white matter damage," *Neurobiology of aging*, vol. 29, no. 10, pp. 1547–1555, 2008.

[64] D. S. Tuch, D. H. Salat, J. J. Wisco, A. K. Zaleta, N. D. Hevelone, and H. D. Rosas, "Choice reaction time performance correlates with diffusion anisotropy in white matter pathways supporting visuospatial attention," *Proceedings of the national academy of sciences*, vol. 102, no. 34, pp. 12212–12217, 2005.

[65] E. Courchesne, H. J. Chisum, J. Townsend, A. Cowles, J. Covington, B. Egaas, M. Harwood, S. Hinds, and G. A. Press, "Normal brain development and aging: quantitative analysis at in vivo mr imaging in healthy volunteers," *Radiology*, vol. 216, no. 3, pp. 672–682, 2000.

[66] A. Pfefferbaum, D. H. Mathalon, E. V. Sullivan, J. M. Rawles, R. B. Zipursky, and K. O. Lim, "A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood," *Archives of neurology*, vol. 51, no. 9, pp. 874–887, 1994.

[67] J. N. Giedd, "Structural magnetic resonance imaging of the adolescent brain," *Annals of the new york academy of sciences*, vol. 1021, no. 1, pp. 77–85, 2004.

[68] A. F. Fotenos, M. A. Mintun, A. Z. Snyder, J. C. Morris, and R. L. Buckner, "Brain volume decline in aging: evidence for a relation between socioeconomic status, preclinical alzheimer disease, and reserve," *Archives of neurology*, vol. 65, no. 1, pp. 113–120, 2008.

[69] L. T. Westlye, K. B. Walhovd, A. M. Dale, A. Bjørnerud, P. Due-Tønnessen, A. Engvig, H. Grydeland, C. K. Tamnes, Y. Østby, and A. M. Fjell, "Life-span changes of the human brain white matter: diffusion tensor imaging (dti) and volumetry," *Cerebral cortex*, vol. 20, no. 9, pp. 2055–2068, 2010.

[70] D. D. Blatter, E. D. Bigler, S. D. Gale, S. C. Johnson, C. V. Anderson, B. M. Burnett, N. Parker, S. Kurth, and S. D. Horn, "Quantitative volumetric analysis of brain mr: normative database spanning 5 decades of life.," *American journal of Neuroradiology*, vol. 16, no. 2, pp. 241–251, 1995.

[71] A. F. Fotenos, A. Snyder, L. Girton, J. Morris, and R. Buckner, "Normative estimates of cross-sectional and longitudinal brain volume decline in aging and ad," *Neurology*, vol. 64, no. 6, pp. 1032–1039, 2005.

[72] T. L. Jernigan, S. L. Archibald, C. Fennema-Notestine, A. C. Gamst, J. C. Stout, J. Bonner, and J. R. Hesselink, "Effects of age on tissues and regions of the cerebrum and cerebellum," *Neurobiology of aging*, vol. 22, no. 4, pp. 581–594, 2001.

[73] T. L. Jernigan and A. C. Gamst, "Changes in volume with age–consistency and interpretation of observed effects.," *Neurobiology of aging*, vol. 26, no. 9, pp. 1271–1274, 2005.

[74] K. B. Walhovd and A. M. Fjell, "White matter volume predicts reaction time instability," *Neuropsychologia*, vol. 45, no. 10, pp. 2277–2284, 2007.

[75] A. R. Luft, M. Skalej, J. B. Schulz, D. Welte, R. Kolb, K. Bürk, T. Klockgether, and K. Voigt, "Patterns of age-related shrinkage in cerebellum and brainstem observed in vivo using three-dimensional mri volumetry," *Cerebral Cortex*, vol. 9, no. 7, pp. 712–721, 1999.

[76] Y. Østby, C. K. Tamnes, A. M. Fjell, L. T. Westlye, P. Due-Tønnessen, and K. B. Walhovd, "Heterogeneity in subcortical brain development: a structural magnetic resonance imaging study of brain maturation from 8 to 30 years," *Journal of Neuroscience*, vol. 29, no. 38, pp. 11772–11782, 2009.

[77] K. B. Walhovd, A. M. Fjell, I. Reinvang, A. Lundervold, A. M. Dale, D. E. Eilertsen, B. T. Quinn, D. Salat, N. Makris, and B. Fischl, "Effects of age on volumes of cortex, white matter and subcortical structures," *Neurobiology of aging*, vol. 26, no. 9, pp. 1261–1270, 2005.

[78] A. M. Fjell, L. T. Westlye, I. Amlien, T. Espeseth, I. Reinvang, N. Raz, I. Agartz, D. H. Salat, D. N. Greve, B. Fischl, *et al.*, "Minute effects of sex on the aging brain: a multisample magnetic resonance imaging study of healthy aging and alzheimer's disease," *Journal of Neuroscience*, vol. 29, no. 27, pp. 8774–8783, 2009.

[79] R. Nesvåg, G. Lawyer, K. Varnäs, A. M. Fjell, K. B. Walhovd, A. Frigessi, E. G. Jönsson, and I. Agartz, "Regional thinning of the cerebral cortex in schizophrenia: effects of diagnosis, age and antipsychotic medication," *Schizophrenia research*, vol. 98, no. 1-3, pp. 16–28, 2008.

[80] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[81] A. M. Brickman, C. Habeck, E. Zarahn, J. Flynn, and Y. Stern, "Structural mri covariance patterns associated with normal aging and neuropsychological functioning," *Neurobiology of aging*, vol. 28, no. 2, pp. 284–295, 2007.

[82] A. M. Fjell, K. B. Walhovd, L. T. Westlye, Y. Østby, C. K. Tamnes, T. L. Jernigan, A. Gamst, and A. M. Dale, "When does brain aging accelerate? dangers of quadratic fits in cross-sectional studies," *Neuroimage*, vol. 50, no. 4, pp. 1376–1383, 2010.

[83] K. W. Schaie, "When does age-related cognitive decline begin?" salthouse again reifies the "cross-sectional fallacy," *Neurobiology of aging*, vol. 30, no. 4, p. 528, 2009.

[84] K. B. Walhovd, L. T. Westlye, I. Amlien, T. Espeseth, I. Reinvang, N. Raz, I. Agartz, D. H. Salat, D. N. Greve, B. Fischl, *et al.*, "Consistent neuroanatomical age-related volume differences across multiple samples," *Neurobiology of aging*, vol. 32, no. 5, pp. 916–932, 2011.

[85] I. J. Deary, A. J. Gow, A. Pattie, and J. M. Starr, "Cohort profile: the lothian birth cohorts of 1921 and 1936," *International journal of epidemiology*, vol. 41, no. 6, pp. 1576–1584, 2012.

[86] B. Dubois, H. H. Feldman, C. Jacova, S. T. DeKosky, P. Barberger-Gateau, J. Cummings, A. Delacourte, D. Galasko, S. Gauthier, G. Jicha, *et al.*, "Research criteria for the diagnosis of alzheimer's disease: revising the nincds–adrda criteria," *The Lancet Neurology*, vol. 6, no. 8, pp. 734–746, 2007.

[87] M. I. Geerlings, C. Jonker, L. M. Bouter, H. J. Adèr, and B. Schmand, "Association between memory complaints and incident alzheimer's disease in elderly people with normal baseline cognition," *American Journal of Psychiatry*, vol. 156, no. 4, pp. 531–537, 1999.

[88] J. Dartigues, C. Fabrigoule, L. Letenneur, H. Amieva, F. Thiessard, and J. Orgogozo, "Epidemiology of memory disorders," *Therapie*, vol. 52, no. 5, pp. 503–506, 1997.

[89] L. K. McEvoy, C. Fennema-Notestine, J. C. Roddey, D. J. Hagler Jr, D. Holland, D. S. Karow, C. J. Pung, J. B. Brewer, and A. M. Dale, "Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment," *Radiology*, vol. 251, no. 1, pp. 195–205, 2009.

[90] M. Jenkinson, M. Pechaud, S. Smith, *et al.*, "BET2: MR-based estimation of brain, skull and scalp surfaces," in *Eleventh annual meeting of the organization for human brain mapping*, vol. 17, p. 167, Toronto., 2005.

[91] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 196–205, 2009.

[92] S. Bauer, T. Fejes, and M. Reyes, "A skull-stripping filter for itk," *Insight Journal*, vol. 2012, 2013.

[93] L. Ibanez, W. Schroeder, L. Ng, and J. Cates, "The itk software guide," *Clifton Park, NY: Kitware*, 2003.

[94] M. W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, and S. M. Smith, "Bayesian analysis of neuroimaging data in FSL," *Neuroimage*, vol. 45, no. 1, pp. S173–S186, 2009.

[95] S. M. Smith, "Fast robust automated brain extraction," *Human brain mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[96] X. Zhang, Y. Feng, W. Chen, X. Li, A. V. Faria, Q. Feng, and S. Mori, "Linear registration of brain mri using knowledge-based multiple intermediator libraries," *Frontiers in neuroscience*, p. 909, 2019.

[97] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[98] N. J. Nilsson, *Principles of artificial intelligence*. Springer Science & Business Media, 1982.

[99] D. B. Lenat and R. V. Guha, *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[100] S. Mor-Yosef, A. Samueloff, B. Modan, D. Navot, and J. G. Schenker, "Ranking the risk factors for cesarean: logistic regression analysis of a nationwide study.," *Obstetrics and gynecology*, vol. 75, no. 6, pp. 944–947, 1990.

[101] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[102] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *Advances in neural information processing systems*, vol. 27, pp. 1799–1807, 2014.

[103] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[104] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure–activity relationships," *Journal of chemical information and modeling*, vol. 55, no. 2, pp. 263–274, 2015.

[105] T. Ciodaro, D. Deva, J. De Seixas, and D. Damazio, "Online particle detection with neural networks based on topological calorimetry information," in *Journal of physics: conference series*, IOP Publishing, 2012.

[106] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[107] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, pp. 2672–2680, 2014.

[108] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *ICLR*, 2017.

[109] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *ICLR*, 2015.

[110] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2794–2802, 2017.

[111] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017.

[112] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *NeurIPS*, pp. 5767–5777, 2017.

[113] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial networks," in *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[114] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

[115] E. L. Denton, S. Chintala, R. Fergus, *et al.*, "Deep generative image models using a laplacian pyramid of adversarial networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[116] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[117] P. J. BURT and E. H. ADELSON, "The laplacian pyramid as a compact image code," *IEEE TRANSACTIONS ON COMMUNICATIONS*, vol. 3, no. 4, 1983.

[118] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018.

[119] C. Wu, L. Herranz, X. Liu, Y. Wang, J. Van de Weijer, and B. Raducanu, "Memory replay gans: learning to generate images from new categories without forgetting," *arXiv preprint arXiv:1809.02058*, 2018.

[120] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *NeurIPS*, 2020.

[121] S. Zhao, Z. Liu, J. Lin, J.-Y. Zhu, and S. Han, "Differentiable augmentation for data-efficient gan training," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[122] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[123] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International conference on machine learning*, pp. 2642–2651, PMLR, 2017.

[124] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: Conditional iterative generation of images in latent space," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4467–4477, 2017.

[125] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

[126] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International Conference on Machine Learning*, pp. 1060–1069, PMLR, 2016.

[127] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7986–7994, 2018.

[128] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," *Advances in neural information processing systems*, vol. 29, pp. 217–225, 2016.

[129] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.

[130] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in *IEEE International Conference on Computer Vision*, 2017.

[131] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International Conference on Machine Learning*, pp. 1857–1865, PMLR, 2017.

[132] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, pp. 2849–2857, 2017.

[133] S. Gurumurthy, R. Kiran Sarvadevabhatla, and R. Venkatesh Babu, "Deligan: Generative adversarial networks for diverse and limited data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 166–174, 2017.

[134] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *International Conference on Learning Representations*, 2014.

[135] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[136] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, pp. 700–708, 2017.

[137] L. Mescheder, S. Nowozin, and A. Geiger, "Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks," in *International Conference on Machine Learning*, pp. 2391–2400, PMLR, 2017.

[138] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*, pp. 1558–1566, PMLR, 2016.

[139] B. Yu, Y. Wang, L. Wang, D. Shen, and L. Zhou, "Medical image synthesis via deep learning," *Deep Learning in Medical Image Analysis*, pp. 23–44, 2020.

[140] Y. Huang, L. Beltrachini, L. Shao, and A. F. Frangi, "Geometry regularized joint dictionary learning for cross-modality image synthesis in magnetic resonance imaging," in *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 118–126, Springer, 2016.

[141] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "Random forest regression for magnetic resonance image synthesis," *Medical image analysis*, vol. 35, pp. 475–488, 2017.

[142] H. Chen, Y. Zhang, W. Zhang, P. Liao, K. Li, J. Zhou, and G. Wang, "Low-dose ct via convolutional neural network," *Biomedical optics express*, vol. 8, no. 2, pp. 679–694, 2017.

[143] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, "Low-dose ct with a residual encoder-decoder convolutional neural network," *IEEE transactions on medical imaging*, vol. 36, no. 12, pp. 2524–2535, 2017.

[144] E. Kang, J. Min, and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction," *Medical physics*, vol. 44, no. 10, pp. e360–e375, 2017.

[145] K. Zeng, H. Zheng, C. Cai, Y. Yang, K. Zhang, and Z. Chen, "Simultaneous single-and multi-contrast super-resolution for brain mri images based on a convolutional neural network," *Computers in biology and medicine*, vol. 99, pp. 133–141, 2018.

[146] L. Xiang, Y. Qiao, D. Nie, L. An, W. Lin, Q. Wang, and D. Shen, "Deep auto-context convolutional neural networks for standard-dose pet image estimation from low-dose pet/mri," *Neurocomputing*, vol. 267, pp. 406–416, 2017.

[147] Y. Wang, B. Yu, L. Wang, C. Zu, D. S. Lalush, W. Lin, X. Wu, J. Zhou, D. Shen, and L. Zhou, "3d conditional generative adversarial networks for high-quality pet image estimation at low dose," *Neuroimage*, vol. 174, pp. 550–562, 2018.

[148] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018.

[149] D. Nie, X. Cao, Y. Gao, L. Wang, and D. Shen, "Estimating ct image from mri data using 3d fully convolutional networks," in *Deep Learning and Data Labeling for Medical Applications*, pp. 170–178, Springer, 2016.

[150] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris, "Multimodal mr synthesis via modality-invariant latent representation," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 803–814, 2017.

[151] X. Han, "Mr-based synthetic ct generation using a deep convolutional neural network method," *Medical physics*, vol. 44, no. 4, pp. 1408–1419, 2017.

[152] A. P. Leynes, J. Yang, F. Wiesinger, S. S. Kaushik, D. D. Shanbhag, Y. Seo, T. A. Hope, and P. E. Larson, "Direct pseudoct generation for pelvis pet/mri attenuation correction using deep convolutional neural networks with multi-parametric mri: zero echo-time and dixon deep pseudoct (zedd-ct)," *Journal of Nuclear Medicine*, 2017.

[153] H. Choi and D. S. Lee, "Generation of structural mr images from amyloid pet: application to mr-less quantification," *Journal of Nuclear Medicine*, vol. 59, no. 7, pp. 1111–1117, 2018.

[154] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris, "Adversarial image synthesis for unpaired multi-modal cardiac data," in *SASHIMI*, 2017.

[155] Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. L. Prince, N. Sugano, and Y. Sato, "Cross-modality image synthesis from unpaired data using cyclegan," in *International workshop on simulation and synthesis in medical imaging*, pp. 31–41, Springer, 2018.

[156] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shapeconsistency generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9242–9251, 2018.

[157] Y. Tsunoda, M. Moribe, H. Orii, H. Kawano, and H. Maeda, "Pseudo-normal image synthesis from chest radiograph database for lung nodule detection," in *Advanced Intelligent Systems*, pp. 147–155, Springer, 2014.

[158] C. Bowles, C. Qin, C. Ledig, R. Guerrero, R. Gunn, A. Hammers, E. Sakka, D. A. Dickie, M. V. Hernández, N. Royle, *et al.*, "Pseudo-healthy image synthesis for white matter lesion segmentation," in *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 87–96, Springer, 2016.

[159] C. Bowles, C. Qin, R. Guerrero, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, "Brain lesion segmentation through image synthesis and outlier detection," *NeuroImage: Clinical*, vol. 16, pp. 643–658, 2017.

[160] X. Chen and E. Konukoglu, "Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders," *Internatinal Conference on Medical Imaging with Deep Learning*, 2018.

[161] S. You, K. C. Tezcan, X. Chen, and E. Konukoglu, "Unsupervised lesion detection via image restoration with a normative prior," in *International Conference on Medical Imaging with Deep Learning*, pp. 540–556, 2019.

[162] N. Pawlowski, M. C. Lee, M. Rajchl, S. McDonagh, E. Ferrante, K. Kamnitsas, S. Cooke, S. Stevenson, A. Khetani, T. Newman, *et al.*, "Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders," *MIDL*, 2018.

[163] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International Conference on Information Processing in Medical Imaging*, pp. 146–157, Springer, 2017.

[164] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain mr images," in *International MICCAI Brainlesion Workshop*, pp. 161–169, Springer, 2018.

[165] H. Uzunova, S. Schultz, H. Handels, and J. Ehrhardt, "Unsupervised pathology detection in medical images using conditional variational autoencoders," *International journal of computer assisted radiology and surgery*, vol. 14, no. 3, pp. 451–461, 2019.

[166] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, "Unsupervised anomaly localization using variational auto-encoders," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 289–297, Springer, 2019.

[167] X. Chen, S. You, K. C. Tezcan, and E. Konukoglu, "Unsupervised lesion detection via image restoration with a normative prior," *Medical image analysis*, vol. 64, p. 101713, 2020.

[168] J. P. Cohen, M. Luck, and S. Honari, "Distribution matching losses can hallucinate features in medical image translation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 529–536, Springer, 2018.

[169] S. Andermatt, A. Horváth, S. Pezold, and P. Cattin, "Pathology segmentation using distributional differences to images of healthy origin," in *International MICCAI Brain-lesion Workshop*, pp. 228–238, Springer, 2018.

[170] E. Vorontsov, P. Molchanov, W. Byeon, S. De Mello, V. Jampani, M.-Y. Liu, S. Kadoury, and J. Kautz, "Boosting segmentation with weak supervision from image-to-image translation," *arXiv preprint arXiv:1904.01636*, 2019.

[171] L. Sun, J. Wang, X. Ding, Y. Huang, and J. Paisley, "An Adversarial Learning Approach to Medical Image Synthesis for Lesion Removal," *arXiv preprint arXiv:1810.10850*, 2018.

[172] C. Chu, A. Zhmoginov, and M. Sandler, "CycleGAN: a Master of Steganography," *NIPS 2017, Workshop on Machine Deception*, 2017.

[173] Y. Zhang, F. Shi, G. Wu, L. Wang, P.-T. Yap, and D. Shen, "Consistent spatial-temporal longitudinal atlas construction for developing infant brains," *TMI*, vol. 35, no. 12, pp. 2568–2577, 2016.

[174] G. Wu, Q. Wang, H. Jia, and D. Shen, "Feature-based groupwise registration by hierarchical anatomical correspondence detection," *Human brain mapping*, vol. 33, no. 2, pp. 253–271, 2012.

[175] W. Huizinga, D. H. Poot, J.-M. Guyader, R. Klaassen, B. F. Coolen, M. van Kranenburg, R. Van Geuns, A. Uitterdijk, M. Polfliet, J. Vandemeulebroucke, *et al.*, "Pca-based groupwise image registration for quantitative mri," *Medical image analysis*, vol. 29, pp. 65–78, 2016.

[176] W. Huizinga, D. H. Poot, G. Roshchupkin, E. E. Bron, M. A. Ikram, M. W. Vernooij, D. Rueckert, W. J. Niessen, and S. Klein, "Modeling the brain morphology distribution in the general aging population," in *Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 9788, p. 97880I, International Society for Optics and Photonics, 2016.

[177] W. Huizinga, D. H. Poot, M. W. Vernooij, G. Roshchupkin, E. Bron, M. A. Ikram, D. Rueckert, W. J. Niessen, S. Klein, A. D. N. Initiative, *et al.*, "A spatio-temporal reference model of the aging brain," *NeuroImage*, vol. 169, pp. 11–22, 2018.

[178] B. C. Davis, P. T. Fletcher, E. Bullitt, and S. Joshi, "Population shape regression from random design data," *IJCV*, vol. 90, no. 2, pp. 255–266, 2010.

[179] A. Serag, P. Aljabar, G. Ball, S. J. Counsell, J. P. Boardman, M. A. Rutherford, A. D. Edwards, J. V. Hajnal, and D. Rueckert, "Construction of a consistent high-definition

spatio-temporal atlas of the developing brain using adaptive kernel regression," *NeuroImage*, vol. 59, no. 3, pp. 2255–2265, 2012.

[180] M. Lorenzi, X. Pennec, G. B. Frisoni, N. Ayache, A. D. N. Initiative, *et al.*, "Disentangling normal aging from alzheimer's disease in structural magnetic resonance images," *Neurobiology of aging*, vol. 36, pp. S42–S52, 2015.

[181] R. Sivera, H. Delingette, M. Lorenzi, X. Pennec, N. Ayache, A. D. N. Initiative, *et al.*, "A model of brain morphological changes related to aging and alzheimer's disease from cross-sectional assessments," *NeuroImage*, vol. 198, pp. 255–270, 2019.

[182] S. Sharma, V. Noblet, F. Rousseau, F. Heitz, L. Rumbach, and J.-P. Armspach, "Evaluation of brain atrophy estimation algorithms using simulated ground-truth data," *Medical image analysis*, vol. 14, no. 3, pp. 373–389, 2010.

[183] M. Modat, I. J. Simpson, M. J. Cardoso, D. M. Cash, N. Toussaint, N. C. Fox, and S. Ourselin, "Simulating neurodegeneration through longitudinal population analysis of structural and diffusion weighted MRI data," in *MICCAI*, pp. 57–64, Springer, 2014.

[184] P. Pieperhoff, M. Südmeyer, L. Hömke, K. Zilles, A. Schnitzler, and K. Amunts, "Detection of structural changes of the human brain in longitudinally acquired mr images by deformation field morphometry: methodological analysis, validation and application," *NeuroImage*, vol. 43, no. 2, pp. 269–287, 2008.

[185] O. Camara, M. Schweiger, R. I. Scahill, W. R. Crum, B. I. Sneller, J. A. Schnabel, G. R. Ridgway, D. M. Cash, D. L. Hill, and N. C. Fox, "Phenomenological model of diffuse global and regional atrophy using finite-element methods," *IEEE transactions on medical imaging*, vol. 25, no. 11, pp. 1417–1430, 2006.

[186] B. Khanal, N. Ayache, and X. Pennec, "Simulating longitudinal brain MRIs with known volume changes and realistic variations in image intensity," *Frontiers in Neuroscience*, vol. 11, p. 132, 2017.

[187] V. Wegmayr, M. Hörold, and J. M. Buhmann, "Generative Aging of Brain MR-Images and Prediction of Alzheimer Progression," in *German Conference on Pattern Recognition*, pp. 247–260, Springer, 2019.

[188] M. F. Rachmadi, M. d. C. Valdés-Hernández, S. Makin, J. M. Wardlaw, and T. Komura, "Predicting the Evolution of White Matter Hyperintensities in Brain MRI using Generative Adversarial Networks and Irregularity Map," *MICCAI*, 2019.

[189] M. F. Rachmadi, M. d. C. Valdés-Hernández, S. Makin, J. Wardlaw, and T. Komura, "Automatic spatial estimation of white matter hyperintensities evolution in brain mri using disease evolution predictor deep neural networks," *Medical Image Analysis*, p. 101712, 2020.

[190] M. Da Silva, C. H. Sudre, K. Garcia, C. Bass, M. J. Cardoso, and E. C. Robinson, "Distinguishing healthy ageing from dementia: a biomechanical simulation of brain atrophy using deep networks," *arXiv preprint arXiv:2108.08214*, 2021.

[191] D. Ravi, D. C. Alexander, and N. P. Oxtoby, "Degenerative Adversarial NeuroImage Nets: Generating Images that Mimic Disease Progression," *MICCAI*, 2019.

[192] D. Ravi, S. B. Blumberg, K. Mengoudi, M. Xu, D. C. Alexander, and N. P. Oxtoby, "Degenerative adversarial neuroimage nets for 4d simulations: Application in longitudinal mri," *arXiv preprint arXiv:1912.01526*, 2019.

[193] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *CVPR*, pp. 5810–5818, 2017.

[194] C. Bowles, R. Gunn, A. Hammers, and D. Rueckert, "Modelling the progression of Alzheimer's disease in MRI using generative adversarial networks," in *Medical Imaging 2018: Image Processing*, 2018.

[195] D. Milana, "Deep generative models for predicting Alzheimer's disease progression from MR data," Master's thesis, Politecnico Di Milano, 2017.

[196] N. Pawlowski, D. Coelho de Castro, and B. Glocker, "Deep structural causal models for tractable counterfactual inference," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[197] Q. Zhao, E. Adeli, N. Honnorat, T. Leng, and K. M. Pohl, "Variational autoencoder for regression: Application to brain aging analysis," *arXiv preprint arXiv:1904.05948*, 2019.

[198] G. Ziegler, R. Dahnke, and C. Gaser, "Models of the aging brain structure and individual decline," *Frontiers in Neuroinformatics*, vol. 6, p. 3, 2012.

[199] H. Yang, D. Huang, Y. Wang, and A. K. Jain, "Learning face age progression: A pyramid architecture of gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 31–39, 2018.

[200] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 2089–2093, IEEE, 2017.

[201] Z. Wang, X. Tang, W. Luo, and S. Gao, "Face aging with identity-preserved conditional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7939–7947, 2018.

[202] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.

[203] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[204] Z. Zhou, H. Cai, S. Rong, Y. Song, K. Ren, W. Zhang, J. Wang, and Y. Yu, "Activation maximization generative adversarial nets," in *International Conference on Learning Representations*, 2018.

[205] A. Borji, "Pros and cons of gan evaluation measures," *Computer Vision and Image Understanding*, vol. 179, pp. 41–65, 2019.

[206] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, pp. 1398–1402, Ieee, 2003.

[207] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*, pp. 2366–2369, IEEE, 2010.

[208] J. P. Pluim, J. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE transactions on medical imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[209] R. P. Woods, S. R. Cherry, and J. C. Mazziotta, "Rapid automated algorithm for aligning and reslicing pet images.," *Journal of computer assisted tomography*, vol. 16, no. 4, pp. 620–633, 1992.

[210] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[211] A. F. Frangi, S. A. Tsaftaris, and J. L. Prince, "Simulation and Synthesis in Medical Imaging," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 673–679, March 2018.

[212] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?," in *International Conference on Learning Representations*, 2019.

[213] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in neural information processing systems*, pp. 2172–2180, 2016.

[214] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Factorised spatial representation learning: Application in semi-supervised myocardial segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, (Cham), pp. 490–498, Springer International Publishing, 2018.

[215] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. C. Courville, "Augmented CycleGAN: Learning many-to-many mappings from unpaired data," in *International Conference on Machine Learning*, 2018.

[216] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Disentangled representation learning in cardiac image analysis," *Medical image analysis*, vol. 58, p. 101535, 2019.

[217] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *European Conference on Computer Vision*, vol. 11207, pp. 179–196, Springer International Publishing, 2018.

[218] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang, "Diverse Image-to-Image Translation via Disentangled Representations," in *European Conference on Computer Vision*, vol. 11205, pp. 36–52, Springer International Publishing, 2018.

[219] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, 2016.

[220] J. R. Taylor, N. Williams, R. Cusack, T. Auer, M. A. Shafto, M. Dixon, L. K. Tyler, R. N. Henson, *et al.*, "The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample," *Neuroimage*, vol. 144, pp. 262–269, 2017.

[221] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[222] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.

[223] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.

[224] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2015.

[225] T. Xia, A. Chartsias, and S. A. Tsaftaris, "Adversarial pseudo healthy synthesis needs pathology factorization," in *International Conference on Medical Imaging with Deep Learning*, pp. 512–526, 2019.

[226] R. B. D'Agostino, "An omnibus test of normality for moderate and large size samples," *Biometrika*, vol. 58, no. 2, pp. 341–348, 1971.

[227] R. D'Agostino and E. S. Pearson, "Tests for departure from normality," *Biometrika*, vol. 50, pp. 613–622, 1973.

[228] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application*, vol. 1. Cambridge university press, 1997.

[229] R. F. Tate, "Correlation between a discrete and a continuous variable. point-biserial correlation," *The Annals of mathematical statistics*, vol. 25, no. 3, pp. 603–607, 1954.

[230] L. Zecca, M. B. Youdim, P. Riederer, J. R. Connor, and R. R. Crichton, "Iron, brain ageing and neurodegenerative disorders," *Nature Reviews Neuroscience*, vol. 5, no. 11, p. 863, 2004.

[231] M. P. Mattson and T. V. Arumugam, "Hallmarks of brain aging: adaptive and pathological modification by metabolic states," *Cell Metabolism*, vol. 27, no. 6, pp. 1176–1199, 2018.

[232] J. H. Cole, R. E. Marioni, S. E. Harris, and I. J. Deary, "Brain age and other bodily 'ages': implications for neuropsychiatry," *Molecular Psychiatry*, vol. 24, no. 2, p. 266, 2019.

[233] C. Jack, R. C. Petersen, Y. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, "Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease," *Neurology*, vol. 51, no. 4, pp. 993–999, 1998.

[234] L. G. Coleman Jr, W. Liu, I. Oguz, M. Styner, and F. T. Crews, "Adolescent binge ethanol treatment alters adult brain regional volumes, cortical extracellular matrix protein and behavioral flexibility," *Pharmacology Biochemistry and Behavior*, vol. 116, pp. 142–151, 2014.

[235] M. Taubert, B. Draganski, A. Anwander, K. Müller, A. Horstmann, A. Villringer, and P. Ragert, "Dynamic properties of human brain structure: learning-related changes in cortical areas and associated fiber connections," *Journal of Neuroscience*, vol. 30, no. 35, pp. 11670–11677, 2010.

[236] Franke, Katja and Ziegler, Gabriel and Klöppel, Stefan and Gaser, Christian and Alzheimer's Disease Neuroimaging Initiative and others, "Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters," *Neuroimage*, vol. 50, no. 3, pp. 883–892, 2010.

[237] J. H. Cole and K. Franke, "Predicting age using neuroimaging: innovative brain ageing biomarkers," *Trends in Neurosciences*, vol. 40, no. 12, pp. 681–690, 2017.

[238] J. H. Cole, S. J. Ritchie, M. E. Bastin, M. V. Hernández, S. M. Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, *et al.*, "Brain age predicts mortality," *Molecular Psychiatry*, 2017.

[239] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[240] B. Jonsson, G. Bjornsdottir, T. Thorgeirsson, L. Ellingsen, G. B. Walters, D. Gudbjartsson, H. Stefansson, K. Stefansson, and M. Ulfarsson, "Deep learning based brain age prediction uncovers associated sequence variants," *bioRxiv*, p. 595801, 2019.

[241] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith, "Accurate brain age prediction with lightweight deep neural networks," *Medical Image Analysis*, vol. 68, p. 101871, 2021.

[242] J. H. Cole, R. Leech, D. J. Sharp, and A. D. N. Initiative, "Prediction of brain age suggests accelerated atrophy after traumatic brain injury," *Annals of neurology*, vol. 77, no. 4, pp. 571–581, 2015.

[243] S. G. Costafreda, I. D. Dinov, Z. Tu, Y. Shi, C.-Y. Liu, I. Kloszewska, P. Mecocci, H. Soininen, M. Tsolaki, B. Vellas, *et al.*, "Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment," *Neuroimage*, vol. 56, no. 1, pp. 212–219, 2011.

[244] L. T. Westlye, K. B. Walhovd, A. M. Dale, T. Espeseth, I. Reinvang, N. Raz, I. Agartz, D. N. Greve, B. Fischl, and A. M. Fjell, "Increased sensitivity to effects of normal aging and alzheimer's disease on cortical thickness by adjustment for local variability in gray/white contrast: a multi-sample mri study," *Neuroimage*, vol. 47, no. 4, pp. 1545–1557, 2009.

[245] F. Farokhian, C. Yang, I. Beheshti, H. Matsuda, and S. Wu, "Age-related gray and white matter changes in normal adult brains," *Aging and disease*, vol. 8, no. 6, p. 899, 2017.

[246] Y. Guo, Z. Zhang, B. Zhou, P. Wang, H. Yao, M. Yuan, N. An, H. Dai, L. Wang, X. Zhang, *et al.*, "Grey-matter volume as a potential feature for the classification of alzheimer's disease and mild cognitive impairment: an exploratory study," *Neuroscience bulletin*, vol. 30, no. 3, pp. 477–489, 2014.

[247] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical image analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[248] L. Fan, H. Li, J. Zhuo, Y. Zhang, J. Wang, L. Chen, Z. Yang, C. Chu, S. Xie, A. R. Laird, *et al.*, "The human brainnetome atlas: a new brain atlas based on connectional architecture," *Cerebral cortex*, vol. 26, no. 8, pp. 3508–3526, 2016.

[249] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *TMI*, vol. 20, no. 1, pp. 45–57, 2001.

[250] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, and R. S. Frackowiak, "A voxel-based morphometric study of ageing in 465 normal adult human brains," *Neuroimage*, vol. 14, no. 1, pp. 21–36, 2001.

[251] D. Mietchen and C. Gaser, "Computational morphometry for detecting changes in brain structure due to development, aging, learning, disease and evolution," *Frontiers in Neuroinformatics*, vol. 3, p. 25, 2009.

[252] J. Oramas, K. Wang, and T. Tuytelaars, "Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks," in *International Conference on Learning Representations*, 2018.

[253] T. Dietterich, "Overfitting and undercomputing in machine learning," *ACM computing surveys (CSUR)*, vol. 27, no. 3, pp. 326–327, 1995.

[254] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[255] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[256] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

[257] C. Gong, D. Wang, M. Li, V. Chandra, and Q. Liu, "Keepaugment: A simple information-preserving data augmentation approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1055–1064, 2021.

[258] X. Wang, H. Chen, H. Xiang, H. Lin, X. Lin, and P.-A. Heng, "Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification," *Medical image analysis*, vol. 70, p. 102010, 2021.

[259] C. Chen, C. Qin, C. Ouyang, S. Wang, H. Qiu, L. Chen, G. Tarroni, W. Bai, and D. Rueckert, "Enhancing mr image segmentation with realistic adversarial data augmentation," *arXiv preprint arXiv:2108.03429*, 2021.

[260] Y. Gao, Z. Tang, M. Zhou, and D. Metaxas, "Enabling data diversity: Efficient automatic augmentation via regularized adversarial training," in *International Conference on Information Processing in Medical Imaging*, pp. 85–97, Springer, 2021.

[261] X. Wang, H. Chen, H. Xiang, H. Lin, X. Lin, and P.-A. Heng, "Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification," *Medical image analysis*, vol. 70, p. 102010, 2021.

[262] R. Hataya, J. Zdenek, K. Yoshizoe, and H. Nakayama, "Meta approach to data augmentation optimization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2574–2583, 2022.

[263] Y. Wang, G. Huang, S. Song, X. Pan, Y. Xia, and C. Wu, "Regularizing deep networks with semantic data augmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[264] X. Zhang, Z. Wang, D. Liu, Q. Lin, and Q. Ling, "Deep adversarial data augmentation for extremely low data regimes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 15–28, 2020.

[265] P. Shamsolmoali, M. Zareapoor, L. Shen, A. H. Sadka, and J. Yang, "Imbalanced data learning by minority class augmentation using capsule adversarial networks," *Neurocomputing*, vol. 459, pp. 481–493, 2021.

[266] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2018.

[267] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "Bagan: Data augmentation with balancing gan," *arXiv preprint arXiv:1803.09655*, 2018.

[268] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, "Gan augmentation: Augmenting training data using generative adversarial networks," *arXiv preprint arXiv:1810.10863*, 2018.

[269] Y. Bin, X. Cao, X. Chen, Y. Ge, Y. Tai, C. Wang, J. Li, F. Huang, C. Gao, and N. Sang, "Adversarial semantic data augmentation for human pose estimation," in *European Conference on Computer Vision*, pp. 606–622, Springer, 2020.

[270] W. Chu, W.-C. Hung, Y.-H. Tsai, D. Cai, and M.-H. Yang, "Weakly-supervised caricature face parsing through domain adaptation," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3282–3286, IEEE, 2019.

[271] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "Metasaug: Meta semantic augmentation for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5212–5221, 2021.

[272] J. Chen and B. Su, "Sample-specific and context-aware augmentation for long tail image classification," 2022.

[273] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2507–2516, 2019.

[274] H. A. Simon, "Spurious correlation: A causal interpretation," *Journal of the American statistical Association*, vol. 49, no. 267, pp. 467–479, 1954.

[275] A. Sinha, K. Ayush, J. Song, B. Uzkent, H. Jin, and S. Ermon, "Negative data augmentation," in *International Conference on Learning Representations*, 2021.

[276] X. Zhang, Q. Wang, J. Zhang, and Z. Zhong, "Adversarial autoaugment," in *International Conference on Learning Representations*, 2019.

[277] A. J. Ratner, H. R. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré, "Learning to compose domain-specific transformations for data augmentation," *Advances in neural information processing systems*, vol. 30, p. 3239, 2017.

[278] J. Ye, Y. Xue, L. R. Long, S. Antani, Z. Xue, K. C. Cheng, and X. Huang, "Synthetic sample selection via reinforcement learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 53–63, Springer, 2020.

[279] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[280] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[281] K. Chaitanya, N. Karani, C. F. Baumgartner, A. Becker, O. Donati, and E. Konukoglu, "Semi-supervised and task-driven data augmentation," in *International conference on information processing in medical imaging*, pp. 29–41, Springer, 2019.

[282] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[283] A. Subbaswamy and S. Saria, "Counterfactual normalization: Proactively addressing dataset shift and improving reliability using causal mechanisms," *arXiv preprint arXiv:1808.03253*, 2018.

[284] M. Temraz and M. T. Keane, "Solving the class imbalance problem using a counterfactual method for data augmentation," *arXiv preprint arXiv:2111.03516*, 2021.

[285] C.-H. Chang, G. A. Adam, and A. Goldenberg, "Towards robust classification model by counterfactual and invariant data generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15212–15221, 2021.

[286] A. Sauer and A. Geiger, "Counterfactual generative networks," in *International Conference on Learning Representations*, 2021.

[287] K. Goel, A. Gu, Y. Li, and C. Ré, "Model patching: Closing the subgroup performance gap with data augmentation," *ICLR*, 2021.

[288] O. Lang, Y. Gandelsman, M. Yarom, Y. Wald, G. Elidan, A. Hassidim, W. T. Freeman, P. Isola, A. Globerson, M. Irani, *et al.*, "Explaining in style: Training a gan to explain a classifier in stylespace," *arXiv preprint arXiv:2104.13369*, 2021.

[289] K. Oh, J. S. Yoon, and H.-I. Suk, "Learn-explain-reinforce: Counterfactual reasoning and its guidance to reinforce an alzheimer's disease diagnosis model," *arXiv preprint arXiv:2108.09451*, 2021.

[290] C. Lu, B. Huang, K. Wang, J. M. Hernández-Lobato, K. Zhang, and B. Schölkopf, "Sample-efficient reinforcement learning via counterfactual-based data augmentation," *arXiv preprint arXiv:2012.09092*, 2020.

[291] S. Dash, V. N. Balasubramanian, and A. Sharma, "Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 915–924, 2022.

[292] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[293] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[294] A. Rahimi, B. Recht, *et al.*, "Random features for large-scale kernel machines.," in *NIPS*, vol. 3, p. 5, Citeseer, 2007.

[295] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *NeurIPS*, 2020.

[296] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European conference on computer vision*, pp. 405–421, Springer, 2020.

[297] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn, "Just train twice: Improving group robustness without training group information," in *International Conference on Machine Learning*, pp. 6781–6792, PMLR, 2021.

[298] V. Feldman and C. Zhang, "What neural networks memorize and why: Discovering the long tail via influence estimation," *arXiv preprint arXiv:2008.03703*, 2020.

[299] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[300] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 289–293, IEEE, 2018.

[301] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks," in *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 1–11, Springer, 2018.

[302] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image synthesis in multi-contrast mri with conditional generative adversarial networks," *IEEE transactions on medical imaging*, 2019.

[303] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[304] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[305] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[306] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[307] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[308] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. S. Torr, and M. Ranzato, "Continual learning with tiny episodic memories.," *CoRR*, vol. abs/1902.10486, 2019.

[309] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *Advances in neural information processing systems*, vol. 30, pp. 6467–6476, 2017.

[310] Z. Chen and B. Liu, "Lifelong machine learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 12, no. 3, pp. 1–207, 2018.

[311] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3366–3375, 2017.

[312] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*, vol. 24, pp. 109–165, Elsevier, 1989.

[313] R. Aljundi, M. Rohrbach, and T. Tuytelaars, "Selfless sequential learning," in *International Conference on Learning Representations*, 2018.

[314] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 532–547, 2018.

[315] A. Gepperth and C. Karaoguz, "A bio-inspired incremental learning architecture for applied perceptual problems," *Cognitive Computation*, vol. 8, no. 5, pp. 924–934, 2016.

[316] A. Robins, "Catastrophic forgetting, rehearsal and pseudorehearsal," *Connection Science*, vol. 7, no. 2, pp. 123–146, 1995.

[317] G. M. van de Ven, H. T. Siegelmann, and A. S. Tolias, "Brain-inspired replay for continual learning with artificial neural networks," *Nature communications*, vol. 11, no. 1, pp. 1–14, 2020.

[318] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang, "An investigation of why overparameterization exacerbates spurious correlations," in *International Conference on Machine Learning*, pp. 8346–8356, PMLR, 2020.

[319] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," *arXiv preprint arXiv:1911.08731*, 2019.

[320] B. Youbi Idrissi, M. Arjovsky, M. Pezeshki, and D. Lopez-Paz, "Simple data balancing achieves competitive worst-group-accuracy," *arXiv e-prints*, pp. arXiv–2110, 2021.

[321] U. Mahmood, R. Shrestha, D. D. B. Bates, L. Mannelli, G. Corrias, Y. E. Erdi, and C. Kanan, "Detecting spurious correlations with sanity tests for artificial intelligence guided radiology systems," *Frontiers in Digital Health*, vol. 3, 2021.

[322] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, pp. 1–10, 2021.

[323] J. M. Wardlaw, M. C. Valdés Hernández, and S. Muñoz-Maniega, "What are white matter hyperintensities made of? relevance to vascular cognitive impairment," *Journal of the American Heart Association*, vol. 4, no. 6, p. e001140, 2015.

[324] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *CVPR*, pp. 1501–1510, 2017.

[325] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[326] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019.

[327] M. C. H. Lee, K. Petersen, N. Pawlowski, B. Glocker, and M. Schaap, "Tetris: Template transformer networks for image segmentation with shape priors," *TMI*, vol. 38, no. 11, pp. 2596–2606, 2019.

[328] G. Jacenkow, A. Chartsias, B. Mohr, and S. A. Tsaftaris, "Conditioning convolutional segmentation architectures with non-imaging data," in *MIDL*, 2019.

[329] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *NeurIPS*, pp. 527–538, 2018.

[330] Z. Li, R. Jiang, and P. Aarabi, "Continuous face aging via self-estimated residual age embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15008–15017, 2021.

[331] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE transactions on image processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[332] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

[333] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

[334] C. Mao, A. Cha, A. Gupta, H. Wang, J. Yang, and C. Vondrick, "Generative interventions for causal learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3947–3956, 2021.

[335] T. Jo, K. Nho, and A. J. Saykin, "Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data," *Frontiers in aging neuroscience*, vol. 11, p. 220, 2019.