Imperial College of Science, Technology and Medicine
Department of Computing

Deep learning for fast and robust medical image reconstruction and analysis

Jo Schlemper

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

September 2019

Declaration of Originality

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

Jo Schlemper

Copyright Declaration

Abstract

Medical imaging is an indispensable component of modern medical research as well as clinical practice. Nevertheless, imaging techniques such as magnetic resonance imaging (MRI) and computational tomography (CT) are costly and are less accessible to the majority of the world. To make medical devices more accessible, affordable and efficient, it is crucial to re-calibrate our current imaging paradigm for smarter imaging. In particular, as medical imaging techniques have highly structured forms in the way they acquire data, they provide us with an opportunity to optimise the imaging techniques holistically by leveraging data. The central theme of this thesis is to explore different opportunities where we can exploit data and deep learning to improve the way we extract information for better, faster and smarter imaging.

This thesis explores three distinct problems. The first problem is the time-consuming nature of dynamic MR data acquisition and reconstruction. We propose deep learning methods for accelerated dynamic MR image reconstruction, resulting in up to 10-fold reduction in imaging time. The second problem is the redundancy in our current imaging pipeline. Traditionally, imaging pipeline treated acquisition, reconstruction and analysis as separate steps. However, we argue that one can approach them holistically and optimise the entire pipeline jointly for a specific target goal. To this end, we propose deep learning approaches for obtaining high fidelity cardiac MR segmentation directly from significantly undersampled data, greatly exceeding the undersampling limit for image reconstruction. The final part of this thesis tackles the problem of interpretability of the deep learning algorithms. We propose attention-models that can implicitly focus on salient regions in an image to improve accuracy for ultrasound scan plane detection and CT segmentation. More crucially, these models can provide explainability, which is a crucial stepping stone for the harmonisation of smart imaging and current clinical practice.

Acknowledgements

I would like to express my gratitude to my supervisor Professor Daniel Rueckert and my second supervisor Professor Jo Hajnal for their guidance throughout the last four years. Without their valuable insights and continuous support, I could not have completed the thesis.

I would like to thank Dr. Ben Glocker and Dr. Bernhard Kainz, the mentors of BioMedIA, for their support not only for the academic work but also making me feel comfortable in the group.

I would like to thank my close collaborators for this thesis: Chen, Guang, Jinming, Jose, Ozan and Wenjia. Without their help and fruitful discussions, the quality of work would have differed significantly.

I would like to thank my collaborators, Anthony, colleagues from Brompton Hospital and St. Thomas Hospital, for their insights into the field as well as sharing data for this thesis.

I would like to thank EPSRC, HIPEDS and SmartHeart project for funding my research over the last four years.

I would like to thank my colleagues from BioMedIA and HIPEDS for making the last four years of my life such a joyful one.

I would like to thank my friends and family for their continuous support and always be there when I needed them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Medical imaging is an indispensable component of modern medical research as well as clinical practice. Despite advances in these imaging techniques, the high demand for the cost as well as the requirement of expert knowledge makes it difficult for them to be used in many scenarios. As such, many parts of the world still do not have sufficient accessibility to these techniques. To make medical devices more accessible, affordable and efficient, it is crucial to reflect upon our current imaging paradigm, reformulate and re-calibrate our problems for smarter imaging.

As each medical imaging technique has a highly structured form in the way it acquires images, one can gather a significant amount of highly correlated data. From such data, one can identify whether there is a redundancy in the information extracted due to sub-optimal imaging procedures. This redundancy in turn provides us with an opportunity to optimise the imaging process holistically. In particular, machine learning algorithms are the techniques which can learn the relationship between different variables in data and optimise the task that one is interested in. The central theme of this thesis is to explore different opportunities where we can exploit implicit or explicit redundancy to improve the way we extract information for better, faster and smarter imaging.

In this section, we take a deeper look into the challenges in current medical imaging in order to motivate the need for smarter imaging protocols. We then highlight the key opportunities that the thesis attempts to address by leveraging advanced information processing techniques

such as deep learning.

1.1 Challenges in modern medical imaging

While medical imaging devices are indispensable for modern medical research, diagnosis, treatment planning and monitoring, many parts of the world still do not have easy access to them. According to world health organization (WHO) report in 2017, while 54% of countries[1] have at least one MRI unit per million population, the numbers are 0% and 30% for low- and lower-middle-income countries respectively, which accounts for the population size of 3.6 billion [Ban16]. Similarly, for CT, the global average is 70%, whereas the number drops to 14% and 60% for the nations with lower income levels [Org17]. The primary reasons of the limited accessibility is due to the vast cost associated with procurement, operation and maintenance, as well as the expert knowledge required to operate them. As such, even if most low-income nations can secure their primary health-care system, there is a huge barrier before these medical imaging devices are incorporated in their national health-care policies.

As of 2016, the leading causes of death for all income groups include cardiovascular diseases, stroke, lung diseases such as chronic obstructive pulmonary disease and cancers, dementia and related disorders (to list a few). While the top leading cause of death for low-income group are more often related to communicable diseases and nutritional conditions, nevertheless, ischaemic heart diseases and stroke still rank within the top five [Org18]. Medical imaging devices can help early diagnosis of these diseases and improve the population health. Therefore, there is an urgent need for the improved dissemination of medical imaging devices across all income levels.

For the better dissemination of medical imaging tools, we need imaging devices that are cheaper, more efficient and easier to use. In order to achieve this goal, we need to reflect upon our current technology stack and rethink how we can optimise the efficiency to maximise the utilisation of these imaging tools.

[1]The report is based on *"Global Atlas of Medical Devices"*, which contained 121 countries from African, American, European, Mediterranean, South-East Asian and Western Pacific regions.

As a research community in medical image analysis, we believe that we can make significant impacts to reduce the operational, maintenance and educational cost by exploiting advanced signal processing techniques and automated/semi-automated methods. In particular, with the emergence of advanced machine learning techniques such as deep learning, we believe that we are in a good position to address these challenges. In the following, we highlight the opportunities that are targeted and tackled in this thesis.

1.1.1 Limitations of acquisition

Medical imaging techniques are inherently complex. Scientists have come up with numerous ways to exploit different branches of physics: nuclear physics, electromagnetism and acoustics. However, each approach has its own limitations. For example, MRI is a non-ionising and non-invasive imaging technique that can offer high-resolution imaging with various contrast mechanisms, yet it is an expensive, slow imaging modality. MRI is dependent on nuclear magnetic resonance (NMR) physics and to obtain a high-quality image, MRI requires a magnet that can produce a strong, homogeneous magnetic field, where the cost of the device increases linearly with the strength of the magnet. Another limitation of MRI is the slow acquisition speed due to its acquisition process, which has a fundamental limit on how fast the data can be acquired due to physiological and hardware constraints. Slow image acquisition limits the availability of the devices, especially for techniques like dynamic MRI, where each clinical examination can last up to 45 minutes [BC17].

X-ray, CT, single photon emission CT (SPECT) and positron emission tomography (PET) are imaging modalities that rely on high-energy electromagnetic radiation. These types of techniques image the attenuation of the emitted radiation doses. They are often cheaper and faster than MRI, however, there is an inherent risk associated with radiation exposure, particularly to the increased rate of cancer incidence [Pea+12; Mat+13; Jou+17].

Another important imaging modality is medical ultrasound (US), which is a cheap and safe imaging modality. US operates by emitting radiofrequency waves and measures the delay in reflected echo to form the image contrast [Sza04]. However, the technology has a much lower

signal-to-noise ratio (SNR) than other imaging modalities, containing many acoustic-based artefacts [Ste+04; Pra+14], which can limit the interpretability and reproducibility of the imaging technique for sonographers who are less skilled [Cha+09].

All imaging techniques therefore have their strengths and weaknesses, and their reliability also depends on applications. Therefore, there is a scope and opportunity to improve the imaging techniques by advancing the algorithms. In particular, most imaging techniques rely on physical principle and do not exploit the enormous amount of data available to optimise the acquisition technique. For each imaging modality, an image is acquired in its respective sensory domain, which is subsequently decoded by inversion techniques. It is expected that huge amounts of redundant data can be exploited to improve the inversion processes, which in return could improve image quality and/or diagnostic output.

1.1.2 Information extraction and processing

In current medical imaging pipelines, there are four major distinct stages: acquisition, reconstruction, analysis (post-processing) and diagnosis. Even though one starts with acquiring an enormous amount of raw data, by the time the data reaches post-processing stage, the information is stripped down to only a few relevant quantities such as presence of abnormality, size and characteristics of the underlying anatomy. In other words, most information is discarded by the time it reaches the diagnostics stage. For example, cardiac MR can be a time-consuming process, yet the radiologist may only be interested in a few quantitative values such as ejection fraction or ventricular mass, before further decisions can be made.

One of the reasons for this inefficiency is that currently, the development of medical images is often focussed on achieving the best image quality, even though ultimately, the most important metric is the diagnostic quality itself. In practice, perfect images are often not necessary and in fact, clinicians and radiologist already live with non-ideal imaging conditions and image artefacts. For example, MRI often contains non-ideal artefacts introduced by either system imperfections or patient motion. In some cases, the images are not usable, however, in many cases, doctors can nevertheless extract required information despite the artefacts. On the

other hand, currently, automated methods such as segmentation tools are often not robust to these image imperfections. For US, image artefacts, such as acoustic shadow, are inherent to the acquisition physics and they can be problematic for subsequent analysis. In often cases, dedicated post-processing techniques are required to specifically correct for them [Men+19].

Therefore, from an application perspective, it is beneficial to approach the medical imaging pipeline holistically, treating the acquisition, reconstruction and analysis as a joint step to be optimised for given target information. We call this paradigm *application-driven imaging*. Reducing the magnitude of acquired data needed to achieve target goals will ultimately lead to more efficient and smarter imaging. This is possible because from information theory perspective, the data already contains all the information needed[2]. Therefore, inefficiency in information processing is another opportunity for advanced data processing algorithms such as machine learning and deep learning. These techniques in principle should be able learn to extract target information more robustly with minimal amount of input, as long as the raw data encapsulates such information.

1.1.3 Interpretation of automated methods

Today, there are increasing numbers of automated/semi-automated approaches that can assist radiologists and clinicians in the current medical imaging diagnostics. For example, this includes automated image segmentation, registration or even image enhancement. These techniques are largely beneficial as they can reduce the burden of tedious manual tasks. As technology evolves, it is expected that one sees an ever-increasing prevalence of these approaches in future.

However, as automated approaches become more pervasive, it will become increasingly important to gain confidence in the mechanics of these algorithms. Having sufficient understandings of how the automated methods work is crucial for dealing with situations such as when they fail. This is especially the case for advanced image reconstruction techniques and holistic approaches that address the aforementioned challenges.

[2]Data processing inequality, the fundamental result in information theory, states that no post-processing of data can increase the amount of information conveyed in it [MM03].

Moreover, understanding the methods will not only provide the confidence for the operators but is also likely to improve the methods themselves. By unravelling how machine learning techniques handle information, it can also facilitate our new understanding of the problems. Machine learning techniques such as neural networks often implicitly find correlation in data, however, less effort has been put to make them interpretable. Addressing this knowledge gap can also help advance the field.

1.2 Objectives and contributions

In this thesis, the aforementioned three problems are targeted as the key opportunities where emerging technologies and algorithms can provide a significant impact and bring us closer to smarter imaging. In particular, at its core, the thesis tackles these challenges by exploiting *data* and advanced algorithms based on *deep learning*. In the last decade, deep learning has made a tremendous impact for individual tasks in computer vision as well as medical imaging. We believe that deep learning, though not limited, is also the key for more advanced imaging approaches in future. The contribution of the thesis is the following:

Deep learning approaches for accelerated dynamic MR data reconstruction

The first challenge identified is the limitation and inefficiency in the current acquisition techniques. For MRI, this is the time-consuming nature of the imaging modality. Inspired by recent advances in deep learning, we propose a framework for reconstructing dynamic sequences of 2D cardiac magnetic resonance (MR) images from undersampled data using deep learning to accelerate the data acquisition process. We show that the proposed methods consistently outperform state-of-the-art compressed sensing methods and are capable of preserving anatomical structures more faithfully up to 11-fold reduction in imaging duration. To this end, we present two approaches: a deep cascade of convolutional neural networks (CNNs) and convolutional recurrent neural network (CRNN). The former exploits the redundancy in data, whereas the

latter exploits the redundancy in both data and the reconstruction process itself for efficient reconstruction.

This contribution has been published in the following:

- **Schlemper, J.**, Caballero, J., Hajnal, J. V., Price, A. N., Rueckert, D. (2017). A deep cascade of convolutional neural networks for MR image reconstruction. Abstract 0643, 25th Annual Meeting and Exhibition International Society of Magnetic Resonance in Medicine, 2017.

- **Schlemper, J.**, Caballero, J., Hajnal, J. V., Price, A., Rueckert, D. (2017, June). A deep cascade of convolutional neural networks for MR image reconstruction. In International Conference on Information Processing in Medical Imaging (pp. 647-658). Springer, Cham.

- **Schlemper, J.**, Caballero, J., Hajnal, J. V., Price, A. N., Rueckert, D. (2017). A deep cascade of convolutional neural networks for dynamic MR image reconstruction. IEEE transactions on Medical Imaging, 37(2), 491-503.

- Qin, C.[†], **Schlemper, J.**[†], Caballero, J., Price, A. N., Hajnal, J. V., Rueckert, D. (2018). Convolutional recurrent neural networks for dynamic MR image reconstruction. IEEE transactions on medical imaging, 38(1), 280-290.[3]

Cardiac MR segmentation from undersampled k-space using deep learning

The second challenge identified is the inefficiency in the way information is extracted and processed. For MRI, reconstruction from undersampled k-space (sensory domain) data enables the accelerated acquisition of MRI but is a challenging problem. However, in many diagnostic scenarios, perfect reconstructions are not necessary as long as the images allow clinical practitioners to extract clinically relevant parameters. For the case of cardiac cine imaging, often quantitative values such as ventricular volume and ejection fraction are used before making further

[3] [†]This is a joint work where each author contributed equally.

decisions for the patient diagnosis. In this thesis, we present a novel deep learning framework for reconstructing such clinical parameters directly from undersampled data, expanding on the idea of application-driven MRI. We propose two deep architectures, an end-to-end synthesis network and a latent feature interpolation network, to predict cardiac segmentation maps from extremely undersampled dynamic MRI data, bypassing the usual image reconstruction stage altogether. We perform a large-scale simulation study show that with the proposed approaches, an accurate estimate of clinical parameters such as ejection fraction can be obtained from very limited amount of raw data per time-frame.

This contribution has been published in the following:

- **Schlemper, J.**, Oktay, O., Bai, W., Castro, D.C., Duan, J., Qin, C., Hajnal, J.V. and Rueckert, D., 2018, September. Cardiac MR segmentation from undersampled k-space using deep latent representation learning. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 259-267). Springer, 2018.

Attention gated networks: learning to leverage salient regions in medical images

The third challenge identified was the need for interpretability of the imaging and analysis techniques as they become more advanced and complex. To this end, we propose a novel *attention gate (AG) model* for medical image analysis that automatically learns to focus on target structures of varying shapes and sizes. Models trained with AGs implicitly learn to suppress irrelevant regions in an input image while highlighting salient features useful for a specific task. This enables us to eliminate the necessity of using explicit external tissue/organ localisation modules when using convolutional neural networks (CNNs). The proposed AG models are evaluated on a variety of tasks, including medical image classification, object localisation and segmentation. For classification, we demonstrate the use case of AGs in scan plane detection for fetal ultrasound screening. We show that the proposed attention mechanism can provide efficient object localisation while improving the overall prediction performance by reducing false positives. For segmentation, the proposed architecture is evaluated on two large 3D CT

abdominal datasets with manual annotations for multiple organs. Moreover, AGs guide the model activations to be focused around salient regions, which provides better insights into how model predictions are made.

This contribution has been published in the following:

- **Schlemper, J.**, Oktay, O., Chen, L., Matthew, J., Knight, C., Kainz, B., Glocker, B., Rueckert, D., Attention-gated networks for improving ultrasound scan plane detection., International Conference on Medical Imaging with Deep Learning, 2018.
- Oktay, O., **Schlemper, J.**, Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., Rueckert, D., Attention U-Net: learning where to look for the pancreas. International Conference on Medical Imaging with Deep Learning, 2018.
- **Schlemper, J.**, Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis, 53, 197-207.

1.3 Thesis overview

This thesis is structured as follows:

- Chapter 2 provides the background for medical imaging acquisition and reconstruction techniques. In particular, it will put a special emphasis on MRI acquisition and reconstruction in detail to provide explanations to why accelerated imaging is desired. The chapter will also briefly cover the fundamentals and the application areas of CT and US, highlighting their respective challenges.
- Chapter 3 provides the background for deep learning in order to provide the reader with sufficient knowledge to understand the methodologies proposed in the thesis. It will also

highlight state-of-the-art approaches that are being used for various tasks in medical image analysis, including classification, segmentation and image reconstruction.

- Chapter 4 and Chapter 5 presents the proposed deep learning approaches for dynamic MR image reconstruction. Chapter 4 covers the approach called a deep cascade of CNN's, whereas Chapter 5 covers the approach based on convolutional recurrent neural network.

- Chapter 6 builds on top of the preceding chapters but takes a step further. In this chapter, we present our proposed approach for direct cardiac segmentation from undersampled cine data, in order to directly obtain clinically relevant parameters from data.

- Chapter 7 addresses the problem of interpretability. In particular, we present the proposed attention-gated networks for improved performance and interpretability, which is important for understanding the neural network-based model which can extract information in an end-to-end fashion.

- Chapter 8 concludes the thesis by summarising the achievements in the thesis, as well as highlight possible and promising future directions.

Chapter 2

Overview of medical imaging techniques

2.1 Introduction

This chapter provides the background information on how images are acquired for each of the medical imaging techniques. In particular, this chapter presents an in-depth description for magnetic resonance imaging (MRI) acquisition and reconstruction, highlighting its fundamental problem of sampling requirement and the limited acquisition speed. This section serves as a motivation for the proposed work on accelerated MR imaging in Chapter 4, Chapter 5 as well as direct segmentation from raw data in Chapter 6.

The second part of the section then briefly covers other imaging modalities, in particular, computational tomography (CT) and medical ultrasound (US). This part provides basic background required for understanding the context of the attention-based methods proposed in Chapter 7.

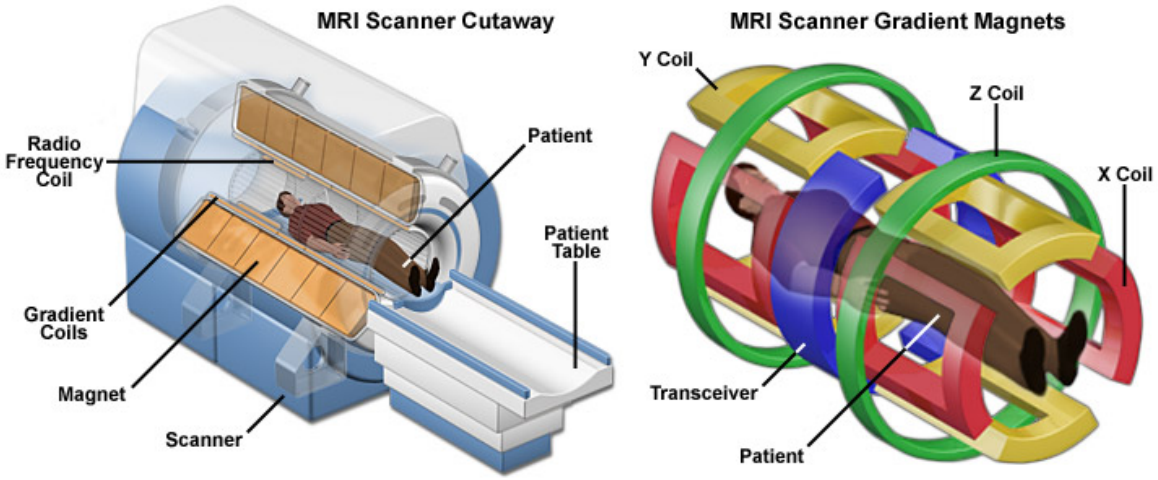

Figure 2.1: (left) The fundamental components of MRI. During an MRI scan, a patient is put into a large cylindrical magnet, which produces a homogeneous magnetic field within the core. A radio-frequency (RF) coil is used to transmit the RF pulse as well as detect the changes in the magnetisation of the body. (right) The configuration of three linear gradient coils. The gradient coils are used to generate linearly varying magnetic field strength along each orthogonal direction. The gradient coils are used to encode the spatial distribution of the magnetisation strength of the underlying object, which is subsequently decoded to form an image (Image courtesy: Maglab [Mag18]).

2.2 Magnetic resonance imaging

The presentation of this section is based on [Nis]. The section will only provide the minimal core concepts required to understand how raw MR data is transformed into an image. For more detail, we refer the readers to the common textbooks [Nis; Haa+99]. The second half surveys the current reconstruction methods for accelerated imaging, including parallel imaging and compressed sensing. The final part introduces the concept of dynamic MR imaging, which is relevant to subsequent chapters.

2.2.1 Imaging principle

Magnetic resonance imaging (MRI) operates by measuring the net magnetisation of hydrogen atoms ^1H abundant in a biological specimen. Hydrogen, being an atom with an odd number of protons, possesses angular momentum called *spin*. Under the influence of a homogeneous magnetic field $\mathbf{B}_0$, two phenomena occur to the spins of hydrogen atoms. Firstly, the spins

align with $\mathbf{B}_0$ in parallel or anti-parallel direction, which results *net magnetisation* $\mathbf{M}$ of the body in the direction of $\mathbf{B}_0$. Secondly, the polarised spins now exhibits resonance at *Larmor frequency* ω:

$$\omega = \gamma B \tag{2.1}$$

where γ is the gyro-magnetic ratio and B is the external magnetic field strength. In MRI, we denote the main static magnetic field as $\mathbf{B_0} = (0, 0, B)$, which by convention is pointing towards z-direction (longitudinal axis). At equilibrium state, the net magnetisation is denoted $\mathbf{M}_0 = (0, 0, M_0)$. In a typical high-field MRI device, the main homogeneous magnetic field $\mathbf{B}_0$ is generated by a cylindrical superconducting magnet – see Fig. 2.1.

2.2.2 Radio-frequency pulse

By applying a polarised *radio-frequency (RF) pulse* $\mathbf{B}_1(t)$ tuned to Larmor frequency, the net magnetisation $\mathbf{M}$ is excited out of its equilibrium state. For the sake of discussion, we assume RF pulse polarised in transverse direction (xy-plane), resulting in the magnetisation $\mathbf{M}$ precessing along xy-plane at Larmor frequency ω. From Faraday's law of induction, a rotating magnetisation vector induces an electromotive force (EMF) proportional to the magnetic field strength. Therefore, by placing a conducting material near the body, the magnitude of the rotating magnetic field can be measured from the induced current. In MRI machine, the component to generate the RF pulse is called RF transmit coil and the component that is used to detect the signal is called RF receiver coil (c.f. Fig. 2.1). The net magnetisation can therefore be measured using the induced current in RF receiver coil that is oriented to detect changes in magnetisation in the xy-plane.

The amplitude and the duration of $\mathbf{B}_1(t)$ determines how far $\mathbf{M}$ is tipped towards the transverse plane. The resulting angle from the longitudinal axis is called *flip angle* θ. When the RF pulse $\mathbf{B}_1$ is switched off, the magnetisation naturally returns to the original equilibrium state $\mathbf{M}_0$. Different body tissues have different rates in which they return to the equilibrium state (c.f.

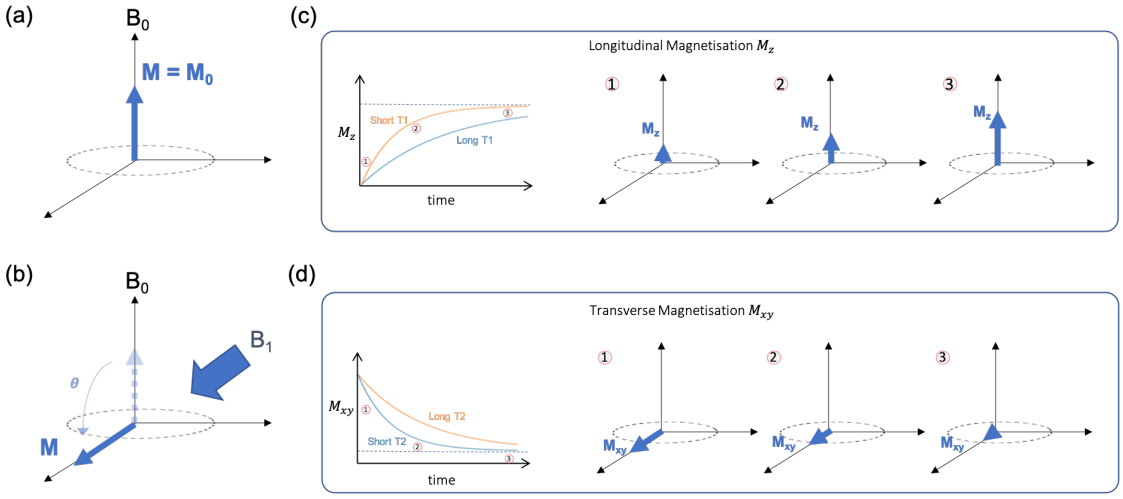

Figure 2.2: (a) The net magnetisation under the influence of $\mathbf{B}_0$. (b) Under the influence of $\mathbf{B}_1(t)$, the spins exhibit resonance and the net magnetisation is tipped towards transverse plane. (c) Once the RF pulse $\mathbf{B}_1$ is switched off, the net magnetisation returns to the equilibrium state. The rate of recovery of the longitudinal component is called T_1 decay, and (d) the rate of signal decay along the transverse direction is called T_2 decay. Note that in reality, $\mathbf{M}$ is precessing about $\mathbf{B}_0$, however, this detail is omitted here for illustration purposes.

Fig. 2.2). Flip angle, T1 and T2 decay are amongst others the parameters that influence the final image contrast. This will be elaborated in Section 2.2.9.

2.2.3 Bloch equation

The above outlines the principle of how the net magnetisation is detected and measured. However, it does not provide us with the ability to distinguish the signals generated from different spatial locations. In order to do so, linear gradient coils in three orthogonal directions are used to encode the spatial distribution of the magnetisation strength into a set of weighted measurements (c.f. Fig. 2.1). The measurements are subsequently decoded by solving a linear equation. The gradient magnetic field is denoted as $\mathbf{G}(t) = (G_x(t), G_y(t), G_z(t))$.

We denote the time varying net magnetisation vector at time t as $\mathbf{M}(t) = (M_x(t), M_y(t), M_z(t))$ and the applied external magnetic field at time t as $\mathbf{B}(t) = (B_x(t), B_y(t), B_z(t))$ as the applied external magnetic field at time t, which contains the influence of the static magnetic field $\mathbf{B}_0$, the RF pulse $\mathbf{B}_1(t)$ and the linear gradient coils $\mathbf{G}(t)$. Under the influence of $\mathbf{B}(t)$, the

trajectory of $\mathbf{M}(t)$ is governed by *Bloch equation*:

$$\frac{\mathrm{d}\mathbf{M}(t)}{\mathrm{dt}} = \mathbf{M}(t) \times \gamma\mathbf{B}(t) - \frac{M_x(t)}{T_2}\boldsymbol{i} - \frac{M_y(t)}{T_2}\boldsymbol{j} - \frac{M_z(t) - M_0}{T_1}\boldsymbol{k} \tag{2.2}$$

The first cross product term in Bloch equation describes the direction in which the torque acts on $\mathbf{M}$ when subjected to the external magnetic field $\mathbf{B}$, which causes the excitation. This encompasses the aforementioned discussion of the effect of $\mathbf{B}_0$ and $\mathbf{B}_1(t)$. The latter three terms describe the change of signal in longitudinal and transverse directions in the absence of $\mathbf{B}_1$. Bloch equation can be extended to study the behaviour of a non-homogeneous object $\mathbf{M}(\mathbf{r}, t)$ under time-varying in-homogeneous magnetic field $\mathbf{B}(\mathbf{r}, t)$, where the solution can be obtained locally at each coordinate $\mathbf{r} \in \mathbb{R}^3$.[1] By acquiring multiple measurements of $\mathbf{M}(\mathbf{r}, t)$ under the influence of time-varying $\mathbf{B}(\mathbf{r}, t)$, one can infer the spatial distribution of $\mathbf{M}$ (i.e. form the image). This process is hereby explained.

Recall that the RF receiver coil measures the sum of all contributions of the magnetic field from all spatial positions under the influence of $\mathbf{B}_0$. This can now be written as:

$$\begin{aligned} s(t) &= \int_{\mathbf{r}\in\mathrm{V}} \mathbf{M}(\mathbf{r}, t) d\mathbf{r} \\ &= \int_{\mathbf{r}\in\mathrm{V}} M^{\circ}(\mathbf{r}) e^{-t/T_2(\mathbf{r})} e^{-\iota\omega_0 t} \exp\left(-\iota\gamma \int_0^t \mathbf{G}(\tau)\cdot\mathbf{r}\mathrm{d}\tau\right) \mathrm{d}\mathbf{r} \end{aligned} \tag{2.3}$$

where the second line of Eq. (2.3) is obtained as a solution of Bloch equation and it consists of four terms and V is the volume under the influence of $\mathbf{B}_0$. Before describing each term, note that as the RF receiver coil only measures the transverse component of the magnetisation, the longitudinal component can be ignored. To describe the two-dimensional transverse component, complex-valued notations are used: i.e., $\mathbf{M}_{xy} = M_x + \iota M_y$, $\mathbf{B}_{xy} = B_x + \iota B_y$. Note that the measurement $s(t)$ is now also complex-valued.

In the second line of Eq. (2.3), $M^{\circ}(\mathbf{r})$ is the initial condition at $t = 0$, i.e., state of magnetisation at position $\mathbf{r}$ right after $\mathbf{B}_1$ is applied. The second term $e^{-t/T_2(\mathbf{r})}$ is the spatially varying T_2 exponential decay effect, which reduces the detectable magnetisation in the transverse direction

[1] In practice, local objects do give microscopic susceptibility effect, called T_2^* effect, which can cause dephasing of the neighbouring tissues, but for modelling purposes this is neglected for simplicity.

after time t. The third term $e^{-\iota\omega_0 t}$ describes the precession of the net magnetisation $\mathbf{M}$ at frequency ω_0 induced by the static magnetic field $\mathbf{B}_0$. The final exponential term describes the effect of gradient magnetic field strength to the precession frequency at position $\mathbf{r}$ after applying the gradient magnetic field over the duration t. For example, applying the gradient magnetic field $\mathbf{G}(t)$ alters the precession frequency at location $\mathbf{r}$ to be $\omega(\mathbf{r}) = \gamma(B_0 + \mathbf{G} \cdot \mathbf{r})$. As such, under the presence of the additional gradient coil, the phase difference accumulates over the duration t. The RF receiver coil receives the aggregation of such position-dependant behaviours as a single measurement as a function of t. The subsequent section describes how multiple of these measurements can be used to form an image.

2.2.4 Signal equation

In order to form an image from the measurement signal $s(t)$, the following is noted. Firstly, the relaxation term $e^{-t/T_2(\mathbf{r})}$ is ignored in the modelling process, assuming the measurement is taken instantaneously. Secondly, one often demodulates the incoming signal, which removes the precession factor $e^{\iota w_0 t}$. Lastly, $M^\circ(\mathbf{r}) = 0$ for all $\mathbf{r} \notin \mathrm{V}$. Therefore, Eq. (2.3) is simplified to:

$$s(\mathbf{k}) = \int_{\mathbf{r}\in\mathbb{R}^3} M^\circ(\mathbf{r}) e^{-\iota 2\pi \mathbf{k}\cdot\mathbf{r}} \mathrm{d}\mathbf{r}, \tag{2.4}$$

where $s(t)$ is re-parameterised as $s(\mathbf{k})$, $\mathbf{k} \in \mathbb{R}^3$, and

$$\mathbf{k}(t) = \frac{\gamma}{2\pi} \int_0^t \mathbf{G}(\tau) \mathrm{d}\tau. \tag{2.5}$$

Eq. (2.4) is called *signal equation*, which shows that the acquired signal $s(\mathbf{k})$ and the MR image $M^\circ(\mathbf{r})$ to be reconstructed have *Fourier relationship*. The idea of the traditional MRI is to form the image M° by acquiring a sufficient set of signals $\{s(\mathbf{k})\}_{\mathbf{k}\in\mathbb{R}^3}$ and perform *inverse Fourier transform*. In the subsequent three sections, we describe how these measurements are obtained and present one instance of the reconstruction methods called Cartesian reconstruction. For this we need the notion of k-space and sampling trajectory.

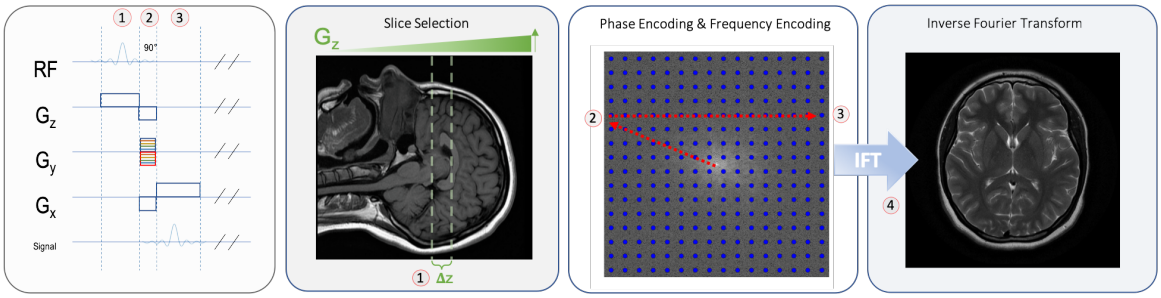

Figure 2.3: The schematic of 2D Cartesian imaging as described in Section 2.2.6. From left: (a) a *sequence diagram*, (b) slice-selection step, (c) phase- and frequency-encoding steps, (d) reconstruction using the inverse Fourier transform. (a) A sequence diagram summarises the information such as when the RF pulse is emitted, when the gradient coils are switched on, for which duration, etc.. The height and the width of the boxes in the sequence diagram show the amplitude and the duration of the selected magnetic field respectively. The gradient is negative if it is below the horizontal line. In 2D Cartesian sampling, The process is repeated for different values of phase-encoding, which is indicated by different colours. (b) The illustration of the slice-selection step, where G_z is applied and the rectangular RF pulse at (1) is used to only excite the specific slice. (c) The phase- and frequency-encoding steps are performed to traverse k-space, where the data is acquired over a rectangular grid. (4) Once enough samples are acquired (the blue points in (c)), the image is reconstructed by taking the inverse Fourier transform of the k-space samples.

2.2.5 k-space

The space in which the signal $s(\mathbf{k}) : \mathbb{R}^3 \to \mathbb{C}$ is acquired is referred to as *k-space* (c.f. Fig. 2.3). k-space encodes the frequency information of the object $M^\circ(\mathbf{r})$ (referred to as an MR image hereafter) and the sample $\mathbf{k}$ corresponds complex sinusoidal basis $e^{-\iota 2\pi \mathbf{k}}$ (c.f. Eq. (2.4)). The measurement $s(\mathbf{k})$ at location $\mathbf{k}$ expresses the amplitude and phase of the basis, which shows the relative contribution of the given basis to an MR image. Therefore, an MR image M° can be seen as a linear combination of sinusoidal waves. MR image acquisition corresponds to a traversal in k-space, measuring $s(\mathbf{k}(t))$ as t is varied over some duration. Note hat k-space can also be defined over two-dimension (2D) for the case when 2D imaging is performed.

2.2.6 Sampling trajectory

In order to reconstruct an image M° from a set of k-space samples, k-space must be sufficiently covered, i.e. a sufficient number of uniform samples in k-space needs to be acquired, before

a linear inversion can be performed. The way in which the k-space samples are acquired is called *sampling trajectory*. Sampling trajectory $\mathbf{k}(t)$ is controlled by applying the gradient coil $\mathbf{G}(t)$. While in theory, an arbitrary sampling pattern can be achieved, in practice, in order to efficiently sample k-space, the sampling trajectory must consider what can be achieved by continuously adjusting gradient coil.

Cartesian sampling

The most common sampling trajectory is 2D *Cartesian sampling* pattern, shown in Fig. 2.3. This sampling pattern reconstructs one 2D image slice at a time and it acquires data on a uniform rectangular k-space grid. For this reason, it is also called rectilinear sampling. In Cartesian sampling, three steps are involved: (1) slice-selection, (2) phase-encoding and (3) frequency-encoding.

(Step 1) slice-selection: In order to perform an imaging of a 2D slice, we select a thin slab of body along z-axis by only exciting $[z_i - \Delta z/2, z_i + \Delta z/2]$ for the i-th slice. This is achieved by switching on $\mathbf{G} = (0, 0, G_z)$ to give a linear variation in ω_0 along z when the RF pulse is emitted. RF pulse also is tuned such that it only excites a specific range of resonant frequencies $[\omega_0 - \gamma G_z z, \omega_0 + \gamma G_z z]$.[2] After slice-selection, the problem is reduced to the imaging of a 2D slice and a pixel in the slice represents the signal contribution along z:

$$m(x, y) = \int_{z-\Delta z/2}^{z-\Delta z/2} M^{\circ}(x, y, \xi) \mathrm{d}\xi \tag{2.6}$$

In addition, 2D k-space is used to describe the acquisition, where the acquired signal is indexed by $s(k_x, k_y)$.

(Step 2) phase-encoding: In phase-encoding step, the k_y location is selected. To do so, $\mathbf{G} = (0, G_y, 0)$ is switched on for t_y such that $k_y = \frac{\gamma}{2\pi} G_y t_y$ (c.f. Eq. (2.5)). Immediately after

[2] The requirement for exciting rectangular frequency profile meant that sinc-like signal needs to be emitted, hence the shape of RF in Fig. 2.3, step 1.

phase-encoding step, the received signal has the form:

$$s(k_x(t), k_y) = \int m(x, y) e^{-\iota 2\pi (k_y y + k_x(t) x)} \mathrm{d}x\mathrm{d}y \tag{2.7}$$

at time t.

(Step 3) frequency-encoding: During frequency-encoding step, $\mathbf{G} = (G_x, 0, 0)$ is switched on, such that $k_x(t) = \frac{\gamma}{2\pi} G_x t$. During this process, the signal is acquired at a sampling rate Δk for the duration of t'.

Two subtleties are noted: firstly, as applying G_z during slice-selection causes the dephasing of precession frequency within the excited slice (c.f. the last term of Eq. (2.3)), a refocussing gradient $-G'_z$ is applied instantaneously to offset this. Secondly, $-G'_x$ is switched on for some time t_x such that the starting point of the sampling is $k_x = -k_{\text{x-max}}$ (e.g. $-k_{x-max} = -\frac{\gamma}{2\pi} G'_x t_x$ and $k_{x-max} = -\frac{\gamma}{2\pi} G_x t'$).

These three steps provide us with multiple measurements of $s(k_y, k_x)$. In particular, in one iteration of the above three steps, one acquires the set of signal $s(k_x, k_y)$ for $k_y = \frac{\gamma}{2\pi} G_y t_y$ and $k_x \in \{-k_{\text{x-max}}, -k_{\text{x-max}} + \Delta k, ..., k_{\text{x-max}} - \Delta k, k_{\text{x-max}}\}$. The three steps are repeated to obtain the samples for different values of k_y's. Once "enough" samples are acquired, the image can be reconstructed. Let N_{PE} be the number of phase encoding performed, and $N_{\text{FE}} = 2k_{\text{x-max}}/\Delta k$. The reconstruction is of $m(x, y)$ on a rectangular grid is performed using inverse Fourier transform as:

$$m(x, y) \approx \sum_{k_y=0}^{N_{\text{PE}}} \sum_{k_x=0}^{N_{\text{FE}}} s(k_x, k_y) e^{\iota 2\pi \left(\frac{k_x x}{N_{\text{FE}}} + \frac{k_y y}{N_{\text{PE}}} \right)} \tag{2.8}$$

where (x, y) and (k_x, k_y) are enumerated based on the rectangular grid indices. The reconstructed MR image $m(x, y)$ is a complex-valued quantity, which contains both the spatial distribution of the magnetisation amplitude $\mathbf{I}(x, y) = |m(x, y)|$ and the phase $\boldsymbol{\theta}(x, y) = \angle m(x, y)$. The magnitude component conveys the anatomical information and therefore $\mathbf{I}$ is used in the most clinical practices. To reconstruct the entire 3D volume, the above process is repeated for

different z_i's.

Non-Cartesian and 3D sampling

2D Cartesian imaging is a common choice in clinical settings as the implementation is simple and fairly robust to system imperfections. However, other sampling trajectories exist, including radial [Kno+11b], spiral [Del+10] and variable density [Kno+11a] trajectories as well as optimised sampling patterns [Laz+19], which in general are referred to as nonuniform/non-Cartesian sampling patterns. Non-Cartesian sampling patterns are attractive due to their motion robustness [Pip99; For+01], however, the reconstruction process is more involved. This is because the k-space measurements no longer align on the rectilinear grid, so one cannot perform discrete Fourier transform for the inversion. In this case, one requires *non-uniform discrete Fourier transform* (NUDFT) [Fes07]. Note that Cartesian sampling can also be performed in 3D: in this case, instead of performing a slice-selection step, the whole volume is excited and it performs phase encoding along two axis ($y-$ and $z-$), followed by frequency encoding. Similarly, any of the non-Cartesian trajectories can be extended to 3D acquisitions.

2.2.7 Sampling requirement

As highlighted above, MR images are reconstructed from a finite set of k-space measurements. In this section, we discuss how many samples are actually required in order to form an image. Such sampling requirement precisely is governed by *Nyquist-Shannon sampling theorem* (NS theorem), which is (informally) stated as:

Theorem 2.1 (Nyquist-Shannon sampling theorem)**.** *A bandlimited continuous-time signal can be sampled and perfectly reconstructed from its samples if the waveform is sampled over twice as fast as its highest frequency component.*

A signal is called *bandlimited by B* if the signal has maximum highest frequency component less than B. NS-theorem says that any sampling rate less than $2B$ will result in aliasing. For

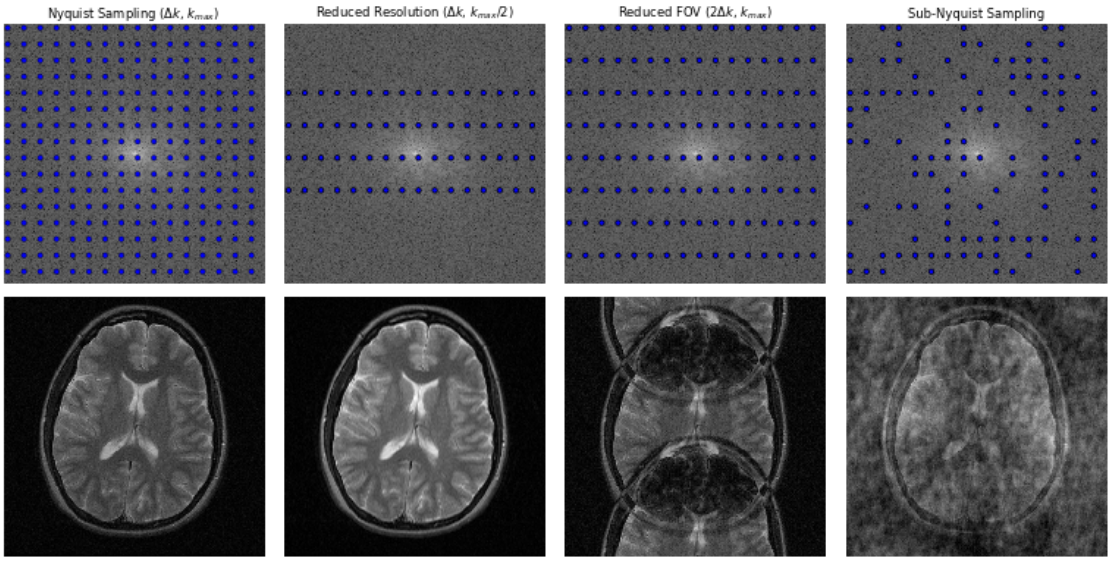

Figure 2.4: The effects of violating Nyquist sampling requirement. From left: fully sampled, reduced $k_{\max}$, increased Δk, and random undersampling. The top row shows the sampling patterns in k-space, whereas the bottom row shows the corresponding resultant aliasing in image domain. In particular, reducing $k_{\max}$ reduces the image resolution, which results in blurry images. When Δk is increased, the FOV is reduced and the replicas are overlapped with each other, causing a wrap-around artefact. When k-space is randomly undersampled, the aliasing appears as incoherent noise. (The brain data shown is taken from [Pau17])

MRI, NS-theorem translates to two fundamental imaging consideration: *field-of-view (FOV)* and *image resolution*. Recall that in Section 2.2.6, we defined two variables: sampling frequency Δk and maximum frequency $k_{\max}$.[3] These quantities are directly related to an FOV and an image resolution.

Field-of-view The sampling frequency Δk affects the FOV inversely:

$$\text{FOV} = \frac{1}{\Delta k}. \tag{2.9}$$

When signal is discretely sampled in k-space at rate Δk, it creates replicas in image domain separated by $1/\Delta k$. Therefore, if FOV is less than the actual underlying object, then the replicas will overlap, creating a *wrap-around* artefact.[4]

Image resolution The maximum frequency $k_{\max}$ obtained in k-space affects the image resolution in the following way:

$$\Delta x = \frac{1}{k_{\max}}, \tag{2.10}$$

where Δx is called the effective resolution of the image, which can be measured in mm/pixel. Effective resolution expresses the how much physical space is occupied by one pixel. Therefore, if $k_{\max}$ is too low, the effective image resolution also becomes low, resulting in blurring of the image. This implies that if high resolution image is required, then the maximum extent of the k-spaced measurements needs to be increased.

Lastly, note that sampling k-space in finite extent also creates *ringing artefact*.[5]

[3] Δk and $k_{\max}$ can differ for each x- and y-axis but the behaviour is the same so we ignore the axis information here.

[4] More precisely, if $s(k)$ is the underlying signal, the discrete sample can be written as $s'(k) = s(k)\text{III}(k/\Delta_k)$, where $\text{III}(k/\Delta_k)$ is a Dirac train separated by k/Δ_k, expressing the sampled locations. Taking the Fourier transform of s' results in a convolution between $\mathcal{F}\{s\}$ and the comb function $\text{III}(\Delta k)$. Convolving with comb function creates a replica separated by Δk.

[5] More precisely, these two artefact can be understood as following: Finite extent in k-space meant that essentially we are multiplying measured signal s with a rectangular function $\Pi_{k_{\max}}$ in frequency domain: $s'' = s(k)\Pi_{k_{\max}}(k)$. Taking a Fourier transform of s'' gives you x convolved with *sinc* function. Convolution with *sinc* function essentially blurs the image by aggregating the local pixel values. In addition, the rippling nature of *sinc* function also gives *ringing artefact* in the image.

The effects of violating NS-theorem are illustrated in Fig. 2.4. Therefore, these two behaviours define the sampling requirements for MR image acquisition: firstly, one needs a sampling rate that is faster than the size of the object such that the object is contained in the FOV. Secondly, one must increase the extent of k-space such that the desired object can be resolved at a sufficient resolution.

2.2.8 Image artefacts

In general, a sampling pattern that does not satisfy NS sampling requirement is referred to as an *undersampling* pattern, and the violation of the sampling requirement results in aliasing. The previous section provided with three types of artefacts: blurring, wrap-around and ringing artefact. More generally, depending on the undersampling trajectory, various artefacts can be observed. For example, random undersampling of k-space can result in complicated noise-like artefact (c.f. Fig. 2.4). Similarly, undersampling based on radial or spiral trajectories can create different coherent artefacts.

Besides from undersampling, there are multiple sources of image artefacts: system imperfections, chemical properties and patient motion. Firstly system imperfections can significantly degrade the image quality. Typically, the acquisition model such as Bloch equation and Fourier relationship assumes an idealised condition (e.g. perfectly homogeneous magnetic field). In the presence of system imperfections such as B_0 inhomogeneity and eddy current, the actual behaviour of the MR physics can deviate from the model, which as a result can cause unexpected distortions in the image. Secondly, chemical properties of the underlying object can also affect the precession frequency, causing further distortions. Lastly, the imaging model assumes that the underlying object is static over the course of imaging. Therefore, patient motion, blood flow, etc., can significantly degrade the image quality. An in-depth overview of image artefact can be found in [Fer+13].

Another source that can affect the image quality is the system noise. In MR imaging devices, thermal noise inevitably affects the measurements. As outlined in Eq. (2.1), the net magnetisation strength is proportional to B_0. Therefore, if B_0 is low, then the signal of the image will

also be low, resulting in images with low signal-to-noise ratio (SNR). This is the primary reason why high-field magnets are preferred in MRI, despite the increase in the cost.

2.2.9 Image contrast

Unlike many other imaging modalities, MRI is a unique technique that allows the generation of images with different contrasts even for the same anatomy. As outlined in Fig. 2.2 as well as in Bloch equation (Eq. (2.2)), the signals from different tissues decay according to their respective T1 and T2 decay parameters. This means that by acquiring data at different time points during the signal decay, one can acquire images with different contrasts. The time between the end of the excitation (the application of $\mathbf{B}_1$) and the signal acquisition (e.g. frequency encoding) is called *echo time* (TE).

As can be seen from Eq. (2.2), it is assumed that each measurement is instantaneous. In practice, due to T2 and T2* effects, the signal decays over time. As such, in order to obtain k-space samples with a consistent contrast, it requires one to repeat the measurement by reapplying RF pulse. The time between successive RF pulses is called *repetition time* (TR).

Based on different combinations of flip angle, TE and TR, images with different contrasts can be created, such as T1-weighted images and T2-weighted images. All above considerations can be summarised in a *sequence diagram*, where an example for Cartesian sampling is shown in Fig. 2.3. A sequence diagram conveys information such as when the RF pulse and the gradient coils are switched on and when the measurements are being made. There are different imaging techniques, and the presented example of Cartesian sampling technique is called *spin-echo* imaging. There are other sequences such as gradient-echo, fast spin-echo, balanced steady-state free-precession (bSSFP) sequence, echo planar imaging, and so on.

2.2.10 Limitation of acquisition speed

So far, we have covered the physics behind MR image acquisition and the imaging considerations in terms of the sampling requirement and image contrast. This section concludes the review

by describing how all of these create an inherent limitation on the acquisition speed of an MRI device. For MRI image acquisition, the total scan duration for a 3D volume can be given as:

$$T_{\text{total}} = \text{ TR } \times \frac{\#\{\text{samples needed per image}\}}{\#\{\text{samples acquired per TR}\}} \times \#\{\text{image slices}\} \tag{2.11}$$

As one can see, the acquisition time required for MRI is directly proportional to the duration of TRs required and the number of image slices to be reconstructed. Unfortunately, this time cannot be easily reduced for the following reasons. Firstly, as outlined in Section 2.2.9, note that TR depends on the desired image contrast. Secondly, there are hardware and physiological constraints that prevent one from traversing k-space faster, limiting the number of samples that can be acquired per TR. This is because, as seen in Eq. (2.4), the speed in which k-space can be traversed is proportional to the duration and the amplitude of the gradient coil $\mathbf{G}$. In order to traverse k-space in a shorter amount of time, the gradient amplitude needs to be increased. However, the speed in which the magnetic field can be increased, called *slew rate*, is limited. In addition, a fast change in magnetic field can cause *peripheral nerve stimulation* (PNS) in the body. As such, there is a fundamental upper bound on the k-space traversal speed for safety reasons.

We further note that in some cases, images may not have sufficient SNR. For example, if an imaging sequence has a long TE, then the net magnetisation along the transverse direction is inherently lower due to T2 decay. In some cases, multiple *signal averaging* is required, which increases the acquisition time by N_{avg}.

Therefore, there is a fundamental limit to MR acquisition speed for reconstructing high-quality images. This has motivated the research community to study methods to reduce the acquisition time. In particular, the main paradigm in which this is achieved is through undersampling of the data, i.e. by reducing the number of samples needed per image. In the previous sections, we highlighted that Nyquist sampling requirement is needed for an alias-free image reconstruction. However, it turns out that certain families of aliasing patterns exhibit coherent or incoherent patterns that can be exploited for image recovery. The following sections now survey accelerated reconstruction techniques.

2.2.11 Accelerated MR image reconstruction

From this section onwards, let $\mathbf{x} \in \mathbb{C}^N$ denote a discrete, combined complex-valued MR image to be reconstructed, represented as a vector with $N = N_x N_y N_z$ where N_x, N_y, and N_z are the width, height and depth of the image. The image can be 2D ($N_z = 1$) or 3D. Let $\mathbf{y} \in \mathbb{C}^M$, $M = n_{\text{sample}}$, represent all the k-space measurements (e.g. $s(\mathbf{k})$) flatted into a vectorised format. Then, the acquired discrete set of samples in k-space can be expressed as:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e} \tag{2.12}$$

where $\mathbf{A} \in \mathbb{C}^{M \times N}$ is called a *forward model*, or a Fourier encoding matrix, and $\mathbf{e}$ is a zero-mean complex Gaussian noise with imaging specific noise variance: $\mathbf{e}_i \sim \mathcal{N}(0, \sigma)$. For example, for 2D Cartesian sampling described in the previous sections, $\mathbf{A}$ represents the sampled Fourier coefficients. In general, $\mathbf{A}$ is generic and it can express Cartesian or nonuniform data acquisition, as well as incorporating multiple weighted measurements for the case of parallel imaging techniques. Our problem is to reconstruct $\mathbf{x}$ from $\mathbf{y}$, formulated as an unconstrained optimization:

$$\underset{\mathbf{x}}{\text{argmin}} \quad \mathcal{R}(\mathbf{x}) + \lambda \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \tag{2.13}$$

where $\mathcal{R}$ is called *regularisation* functional on $\mathbf{x}$ and λ is a hyper-parameter often associated with the noise level of the input. The regularisation functional $\mathcal{R}$ encodes our prior knowledge about the image to be reconstructed. The form of $\mathcal{R}$ is explained in subsequent sections.

Parallel MRI

In parallel imaging, data is acquired using N_{coil} receiver coils, where each receiver coil is more sensitive to signals generated in the proximity of the coil, as shown in Fig. 2.5. Mathematically, the coil sensitivity maps can be expressed by spatially varying weights $\mathbf{S}^{(i)} \in \mathbb{C}^{N \times N}$ for the coil i, and the image seen by the coil i can be expressed as $\mathbf{S}^{(i)}\mathbf{x}$. Then, for each receiver coil, the

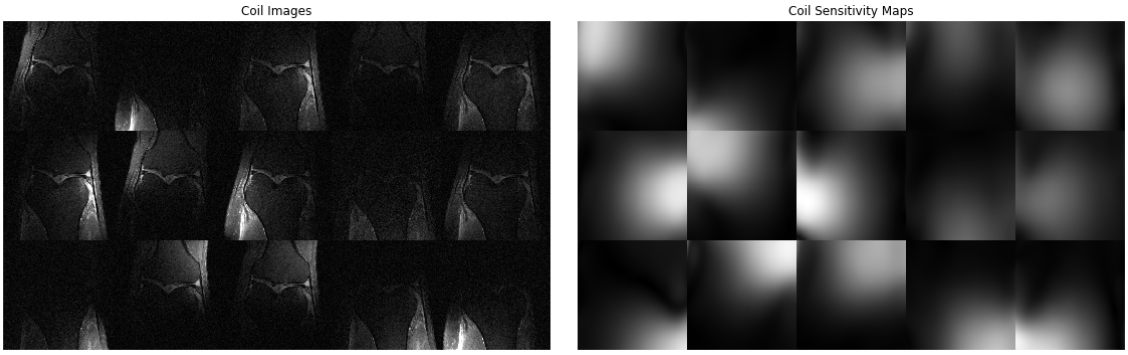

Figure 2.5: Parallel Imaging uses multiple receiver coils. The use of multiple weighted images can be seen as a "spatial encoding" – they provide explicity redundancy in . data to enable accelerated MR image reconstruction. (The knee image shown here is from fastMRI challenge dataset [Zbo+18])

measurement data will have the form:

$$\mathbf{y}^{(i)} = \mathbf{F}\mathbf{S}^{(i)}\mathbf{x} + \mathbf{e}^{(i)}. \tag{2.14}$$

Provided that the coil sensitivity is known, the coil images can be combined into a single image in an SNR optimal way pixel-wise [Roe+90]. Without loss of generality, assume that for each pixel p, the coil is normalised: i.e. $\sum_{i=1}^{N_{\text{coil}}} \mathbf{S}_p^{(j)H}\mathbf{S}_p^{(j)} = 1$. Then, for each pixel p, the coil combination is given by:

$$\mathbf{x}_p = \sum_{i=1}^{N_{\text{coil}}} \mathbf{S}_p^{(i)H}\mathbf{x}_p^{(j)} \tag{2.15}$$

The idea of accelerated reconstruction started with the observation that the images from multiple receiver coils can provide explicit redundancy in data [SM97; Jak+98; Hei+01]. In order to accelerate imaging, one can perform a regular undersampling with undersampling factor R. For example, for $R = 2$, every other line is skipped along phase-encoding direction. As it was discussed in Section 2.2.7, this reduces the FOV by a factor of 2, causing a wrap-around artefact. However, because there are multiple coil images with different coil sensitivity maps, the overlapped pixel contribution can be inverted from the redundant information. This approach is called *sensitivity encoding* (SENSE).

More generally, consider a forward matrix $\mathbf{A}$ of the form:

$$\mathbf{A}_{(\gamma,\kappa),\rho} = e^{\iota \mathbf{k}_\kappa \cdot \mathbf{r}_\rho} s_\gamma(\mathbf{r}_\rho) \tag{2.16}$$

This is a forward matrix for parallel reconstruction setting, where $\mathbf{A} \in \mathbb{C}^{N_{\text{coil}}M \times N}$. γ, κ, ρ index coil, k-space and pixel position respectively. If the number of samples MN_{coil} is greater than number of image pixels N, then the system of equation $\mathbf{Ax} = \mathbf{y}$ is over-determined and it can be solved using pseudo-inverse:

$$\mathbf{x} \approx \left(\mathbf{A}^H \mathbf{A}\right)^{-1} \mathbf{A}^H \mathbf{y} \tag{2.17}$$

For the case of regular Cartesian undersampling (e.g. skipping every R phase encoding steps), Eq. (2.17) reduces to solving the pseudo-inverse using only the pixels that are overlapped due to the reduced FOV. In particular, with no undersampling factor, the solution reduces to Eq. (2.15). In case where undersampling pattern is non-Cartesian, then the full matrix inversion is required. As $\mathbf{A}$ is a large matrix, the direct calculation of the pseudo-inverse is prohibitive. CG-SENSE [Pru+01] is an efficient algorithm which solves the inversion iteratively.

Another well-known approach is GRAPPA. The idea of GRAPPA is to exploit the redundancy in k-space representations of the coil images to synthesise the missing k-space points from their neighbouring points. This can be done due to following two observations. Firstly, since a coil sensitivity map has a smoothly varying profile in image domain, the corresponding frequency domain representation has a small support around zero, only containing low frequencies. Secondly, since an element-wise multiplication is a convolution in frequency domain, this meant that each k-space representation of the coil image is simply the convolution between the k-space representation of $\mathbf{x}$ and a small kernel. Therefore, one should be able to learn a transition-invariant deconvolution kernel to "undo" the effect of the coil sensitivity to each coil image. This deconvolution kernel is estimated from a small set of *auto-calibration lines*, which is a fully-sampled low-frequency region.

SENSE and GRAPPA achieve similar acceleration rates but they have slightly different arte-

facts due the different approximation techniques being employed. Nevertheless, SENSE and GRAPPA are are related and this is generalised in E-SPIRIT [Uec+14].

g-factor When parallel imaging for accelerated MR image reconstruction is performed, the SNR of the reconstructed image decreases due to two factors. Firstly, since the amount of data aquired is reduced by a factor R, the SNR is reduced by $\sqrt{R}$. Secondly, when the acceleration factor is increased in SENSE/GRAPPA, the FOV is reduced, causing the underlying object to wrap around in the limited FOV. However, noise also wraps around, where the source cannot be distinguished. As a result, one observes spatially variant noise amplification. This is called g-factor [AVT14]. The resulting SNR of the accelerated image is therefore given by:

$$\mathrm{SNR}_R = \frac{\mathrm{SNR}_{\mathrm{full}}}{g\sqrt{R}}. \tag{2.18}$$

where $\mathrm{SNR}_{\mathrm{full}}$ is the SNR of the fully-sampled image. g-factor for GRAPPA is similar but is given by the interpolation kernel. Due to such SNR consideration, there is a fundamental limit to how much the image can be accelerated using parallel imaging. In order to overcome this barrier, powerful nonlinear reconstruction is needed that can infer the underlying information, overcoming the effect of noise.

For more in-depth survey of parallel MRI, we refer the readers to [Des+12; LN07].

Compressed sensing MRI

Compressed sensing is a complementary technique to perform accelerated MR image reconstruction. Whereas parallel imaging relies on explicit redundancy from multiple coil images, compressed sensing relies on implicit redundancy in data. This technique is now surveyed.

In Eq. (2.12), if $M << N$, the equation is under-determined and ill-posed. In this case, there are infinitely many solutions. In order to obtain a unique solution, one must supplement the inverse problem with additional constraints using regularisation $\mathcal{R}$, as seen in Eq. (2.13). Compressed sensing (CS) [Don+06; EK12; Can08] is a theoretically sound framework which

can recover a unique solution for such ill-posed system, exploiting the *sparsity* of the signal. In general, there are three requirements:

- a sparsity of the signal,
- incoherent sampling measurements and
- a nonlinear optimisation algorithm.

Firstly, for CS to apply, the measured image must have a sparse representation. Data is called *sparse* if the data points can be represented by a linear combination of only a few basis points. Natural images and medical images indeed admit sparse representations, which can be demonstrated from JPEG compression.

The second requirement is that the acquisition matrix needs to be *incoherent*. If the measurements are coherent, then the aliasing is structured and there is no way of distinguishing whether such structure is an aliasing or an actual part of the image that can be sparsely represented. It turns out that the random acquisition in k-space is indeed sufficiently incoherent, generating noise-like artefact in image domain (c.f. Fig. 2.4).

Once above two conditions are met, then the signal can be reconstructed by solving the following optimisation problem:

$$\begin{aligned} \underset{\mathbf{x}}{\text{argmin}} \quad & \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{\Phi}\mathbf{x}\|_0 \leq k \end{aligned} \tag{2.19}$$

where $\mathbf{\Phi}$ is a sparsifying transform of $\mathbf{x}$ and $\|.\|_0$ is an ℓ_0 norm, which counts the number of non-zero entries and k is the maximum number of coefficients that are allowed in the sparse representation. In practice, solving for such sparse solution is an *NP-hard* problem, as selecting k non-zero entries is combinatorial in nature. However, it turns out that the *convex* relaxation of Eq. (2.19) robustly converges to the same sparse solution:

$$\underset{\boldsymbol{\gamma}}{\text{argmin}} \quad \|\boldsymbol{\gamma}\|_1 + \lambda\|\mathbf{A}\mathbf{\Phi}^H\boldsymbol{\gamma} - \mathbf{y}\|_2^2 \tag{2.20}$$

where $\mathbf{x} = \mathbf{\Phi}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is the corresponding sparse representation of $\mathbf{x}$. The Eq. (2.20) is called *LASSO*. The attractiveness of CS is that there is a theoretical guarantee on the number of measurements M required to perfectly reconstruct $\mathbf{x}$ of the dimensionality N that is k-sparse in $\mathbf{\Phi}$.

The analysis of CS almost directly translates to the case of MRI reconstruction problem. As the acquisition time is proportional to the number of samples required in k-space, CS offers a great potential for accelerating MR image reconstruction. In particular, Eq. (2.20) has the form of Eq. (2.13), where regularisation $\mathcal{R}$ corresponds to ℓ_1 norm in sparse domain. Therefore it can be efficiently solved using convex or nonlinear optimisation algorithms. The first successful application of CS is by Lustig *et al.* in [Lus+05], where the authors have showed that 2 to 4 fold acceleration can be achieved for brain images, and up to 20-fold acceleration for MR angiography applications.

Since their seminal work was published in 2005, CS-MRI has become an active area of research. As aforementioned, for CS, the number of required samples for the near perfect reconstruction is directly related to the sparsity of the signal. Therefore, the natural question in CS is the optimal sparse representation for a specific application area. For CS-MRI, many sparse representations for medical imaging applications have been proposed, leading to advanced regularisation terms such as total generalised variation (TGV) [Kno+11b] and structured sparsity [CH12]. One notable approach is *dictionary learning*, which is an approach that tries to jointly optimise for the reconstruction and the optimal sparse basis. By making the sparse representation adaptive, further acceleration can be achieved [RB11; Cab+14b]. CS-MRI has successfully been integrated with parallel imaging, such as ℓ_1-SPIRIT [Mur+12].

More recently, *low-rank* approaches have been studied in the context of CS-MRI. Low rank approaches can be seen as a generalisation of compressed sensing to multi-dimensional data. In particular, by assuming the low-rankness (e.g. sparsity in spectral domain), one can perform matrix completion from limited number of samples. For parallel imaging, block-structure Hankel matrix can be defined for the acquired multi-coil data, which is low-rank. This can be used as a constraint to recover parallel imaging, such as SAKE, LORAKS, p-LORAKS [Shi+14;

Hal13; HZ16]. The framework is generalised to annihilating filter-based low rank Hankel matrix approach (ALOHA) approach [JLY16].

CS-MRI techniques are currently gradually being deployed in clinical settings and it is still an active area of research [Vas+11; Hea19]. However, CS-MRI has several limitations. Firstly, it is still an open question what the optimal sparse representation is for a particular application. For example, for MR angiography where the images are sparse in image gradient domain, total variation (TV) provides a sufficient regularisation. However, for Muscoskeletal imaging where texture of the image is important, TV can yield undesirably blocky artefacts. Secondly, even with the suitable regularisation, the quality of the reconstruction still largely depends on the weighting term λ. As can be seen from Eq. (2.12), λ term balances between the data fidelity and regularisation term. The optimal lambda is application specific, which depends on the suitability of the regularisation term and the noise level present in the data. Therefore, typically, a hyper-parameter search is required. Some methods exist to automatically estimate the appropriate values of lambda from the noise-level of the data. However, it is still unclear if this approach can necessarily provide a good perceptual quality independent of the applications. Lastly, CS-MRI techniques often only consider a single image recovery. This motivates more powerful approach such as bi-level optimisation and machine learning approaches, which are surveyed in the next chapter.

2.2.12 Dynamic MRI

In dynamic MRI, the aim is to characterise the anatomies in motion. The application area includes cardiovascular MR (CMR) and/or perfusion imaging such as late-Gadolinium enhanced imaging (LGE-MRI). In this section, we focus our attention on CMR due to the relevance for the subsequent chapters, but the principle applies for other dynamic imaging applications.

CMR is a gold-standard clinical tool for the evaluation of cardiac morphology and function [CBR13]. In cardiac *cine* imaging, an accurate tracking of the cardiac phases is required. A cardiac cycle can be monitored using electrocardiogram (ECG) and is typically characterised by *QRS-complex*. There are two dominant cardiac phases, which are called *systolic* and *diastolic*

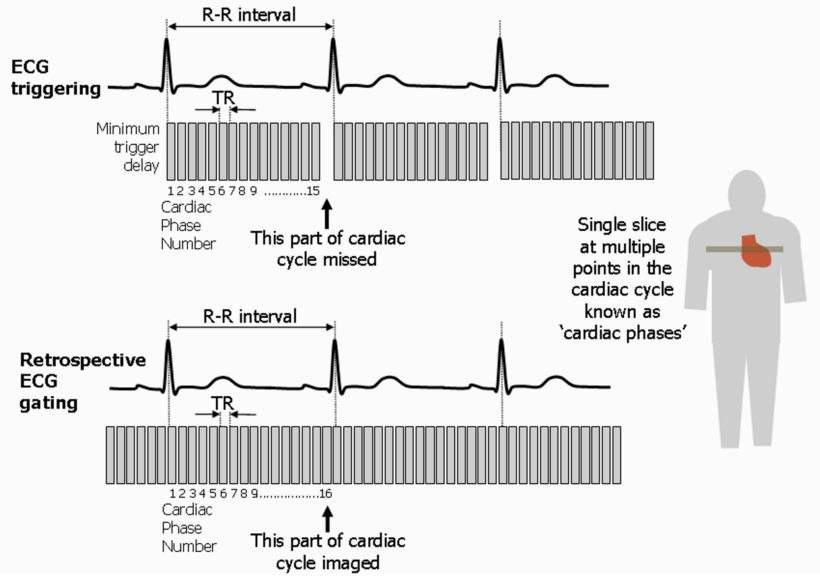

Figure 2.6: ECG triggering and retrospective ECG gating for dynamic MRI. ECG triggering starts the acquisition once the 'R' wave is detected. On the other hand, retrospective ECG gating continuously acquire data over multiple cardiac phases, where the data is retrospectively binned to create the images. (Image courtesy: [Rid10])

phases. In systolic phase, the myocardium contracts to perform ejection of blood. In diastolic phase, the myocardium is relaxed and the ventricular volume increases. While ECG can be used to monitor the cardiac phases, cine-imaging can enable accurate depiction and quantification of the anatomy.

While the underlying imaging principle of dynamic imaging is identical to that of static imaging, it has several unique challenges. Firstly, it is not possible to obtain an image with both high temporal and spatial resolution from a single cardiac cycle as the duration of one cardiac cycle is too short. This forces one to use imaging techniques with TR's that are much shorter than one cardiac cycle. This restricts the imaging protocols to be gradient echo techniques, including refocussed gradient echo (GRE), bSSFP or echo-planar imaging (EPI). Nevertheless, often the k-space signals are acquired over several cardiac cycles, which are retrospectively combined into an image sequence of a single heart-beat.

For this binning to work, there is a need for the synchronisation of the imaging protocol and the cardiac cycles. There are two ways in which this can be achieved: *ECG triggering* and *retrospective ECG gating*. In ECG triggering, the 'R' wave of the cardiac phase is detected,

which triggers the imaging protocol. This enables one to acquire the image at a specified cardiac phase, specified by trigger delay. The alternative is to use retrospective ECG gating. In this latter approach, the data is continuously acquired during the cardiac cycle while the patient's cardiac cycle is monitored using ECG. At the end of data acquisition, the data is retrospectively binned into correct cardiac phases. While the latter is typically more efficient, if the patient has arrhythmia (high beat-to-beat duration variability), then cine-imaging itself is challenging and in this case ECG triggering may be more suitable if only a static image of systolic phase needs to be acquired.

The second problem is motion. In addition to cardiac motion, there is also a respiratory motion, which can introduce serious image degradation for CMR if not handled appropriately. There are three ways in which this can be addressed: breath-hold acquisition, respiratory gating and respiratory binning techniques. The most common approach in current CMR protocols is the breath-hold acquisition, in which the patient is asked to hold the breath during signal acquisition [Sta+16]. While this is the simplest and efficient for cine imaging, it also induces significant demand on patients and is difficult for uncooperative patients. Respiratory gating and respiratory binning are used when imaging is too time-consuming or when breath-hold is unsuitable, such as 3D imaging [Bus+19] and cardiac magnetic resonance fingerprinting [Cru+19]. For more in-depth survey of CMR techniques, refer to [Rid10; BRR12].

Therefore, accelerating CMR protocol is of high interest as reducing the acquistion time not only reduces the demand on patient but also improves the image quality by reducing the chance for motion artefact to influence. Indeed, CMR is suitable for CS-based accelerated reconstruction techniques because CMR image sequence exhibits high spatio-temporal correlation between each time frame, which therefore admits sparse representation in image and time domains. To this end, k-t SENSE and k-t BLAST were proposed [TBP03], which exploits the spatio-temporal correlation to perform parallel imaging techniques. This was later improved by *k-t* FOCUSS [JYK07] to include compressed sensing. For CMR, cardiac images are static for most of the anatomy except for the cardiac anatomy. This characteristics can be precisely captured by combining compressed sensing and low-rank approaches. In particular, the static part is reconstructed using low-rank constraint and the remaining moving anatomy

are reconstructed using sparsity constraint. Well-known implementations of these approaches are k-t SLR [Lin+11], L+S [OCS15] and STORM [PJ15]. Dictionary learning method has also been extended for dynamic MRI [Cab+14b]. For free-breathing approaches, there are notable methods such as XD-GRASP [Fen+16].

2.3 Computational tomography

Computational Tomography (CT) in essence reconstructs image from the measurements of projections, which is an extension of an X-ray technique proposed by Hounsfield in 1970 [Hou73; SB95]. Modern CT acquires data by passing multiple X-rays along patient. The attenuation of X-ray depends on tissue density, which is called attenuation coefficient. The spatially varying distribution of attenuation coefficient forms the basis of CT images.

In modern 2D CT imaging, the emitted X-rays form a fan-beam geometry. The emission source and the radially laid out detectors are rotated to obtain measurements from all angles. The complete set of line integrals is called *Radon transform*, which uniquely determines the object $\mathbf{x}$. A set of Radon transformed data is called a *sinogram*. CT therefore defines an inverse problem similar to MRI: the forward matrix $\mathbf{A}$ expresses the different line integrals parameterised by the fan-beam angles [SP15].

Fourier slice theorem tells that the 1D Fourier transform of Radon projections of an object and the parallel line in 2D Fourier transform of an object are the same. In particular, the underlying object is uniquely determined by Radon transform, provided that there are sufficient measurements. In practice, the reconstruction are made from finite set of samples, as in MRI. There are many reconstruction techniques for CT imaging: analytical methods and iterative methods. For analytical method, filtered back projection (FBP) is the most common approach, which weighs each projection by Radon kernels. Different kernel provides trade-off between noisiness and sharpness. On the other hand, iterative reconstructions solves the regularised inverse problem as seen in Eq. (2.13). Iterative reconstructions are attractive as they enable one to incorporate the prior knowledge about the underlying anatomy, which enables compressed

sensing based reconstructions. While iterative approaches are more computationally expensive than FBP, owing to the advances in hardware, the iterative approaches are gaining popularity.

Modern CT enables one to acquire very high resolution images (e.g. 1024×1024 pixels for a 2D slices). The main concern, however, is the risk of radiation exposure [Jou+17; Pea+12; Mat+13]. The added lifetime risk of developing cancer by a single abdominal CT is estimated to be one in 2,000 [Sch16]. To this end, a significant amount of effort has been put to reduce the dosage. The major classes of approaches for reduced-dose imaging are low-dose CT [Tea11; Kan+96; Nai+90], limited angle CT [HW83; Dav83; DB98], and interior tomography [WY13]. A number of compressed sensing based methods have been utilised for these problems [SP08; Ham+13; MF15; McC+16] and more recently, deep learning techniques (surveyed in Chapter 3). The major challenge is that the image quality is inherently linked to the dosage used. In essence, the variance of measurement increases if the dose is reduced, yielding a noisier image. Therefore, obtaining good image quality from reduced dosage is still an ongoing research.

CT has a wide range of application, especially when a high resolution image is required where the object may be susceptible to motion, or in a time constraint environment such as emergency room [Nov+99; GR08]. This includes lung CT [Kan+96] , CT cardiac screening [Nik+04; Hau+09], high resolution breast CT [Boo+01] and whole body CT [Hub+09].

2.4 Ultrasound imaging

In this section, we provide a brief overview of medical ultrasound. An excellent overview of the imaging technique can be found in [CSJ11; Fin92; WTF92].

Diagnostic ultrasound (US), or ultrasonography, is a noninvasive imaging technique used to visualise internal body structures as well as motion using ultrasound physics, introduced in 1960's for medical use [Org+98]. US consists of a transducer, transmitter pulse generator, amplifier, analog-to-digital converter and computer processing unit which post-processes the image and display. B-mode US is the most fundamental imaging technique, where a linear array of transducers are used to simultaneously emit a radio-frequency wave to scan a two-

dimensional plane. When the emmitted pulse hits an object (or change in medium), reflection occurs. The ultrasound in essence visualises an internal organ by measuring the delay in echo arrival [Fin92; WTF92].

Many extensions of B-mode US exists. Firstly the B-mode US can be extended to image the underlying object in real-time. In M-mode US, a rapid sequence of B-mode scans are performed. A three-dimensional imaging is also possible by performing successive two-dimensional scans of adjacent planes, or by using two-dimensional array of transducers. Another notable variation of US is doppler-mode US. In this technique, the doppler shift cause by the reflection from moving objects is measured, which allows one to compute, for example, flow information [CSJ11].

As ultrasound is cheap and safe to use, it is an indispensible part of modern diagnostic imaging. However, the acquisition physics yields a fundamental limitation to the image quality and the produced images contain many artefacts. For example, an object that is either reflective or attenuating can disrupt the echo, creating shadows in the image where no signal can be detected. Another source of artefact is a circular object, which can cause a refraction that results in added noise in the images. US is often employed in a clinical settings where a real-time imaging is required. As such, it is also challenging to employ computationally expensive reconstruction techniques for post-processing the images to mitigate these image artefacts. Nevertheless, compressed sensing approach have been proposed to improve image quality [Ach+10; Qui+10; Wag+12; LPF13]. More recently deep learning methods are also investigated to adaptively postprocess the image [KHY19].

Today, medical US has a wide impact for clinical practice both in developed and developing world. The application area includes echocardiography [NQ14], obstetrics and gynaecology [Cal88; Cam13] and disease monitoring in developing countries [Cal88; Met91] to list a few. However, in the views of ultrasonography's widespread application, the training of the sonographers as well as maintaining the qualitative diagnosis has become an important challenge. A poor use of the device can result in misdiagnosis and errors in interpretation, which can be harmful in some cases [LL91].

2.5 Summary

In this section, we highlighted some of the challenges in each medical imaging devices. In particular, we delved in to MRI acquisition, which serves as a key example how imaging approach creates fundamental limitation to the acquisition. We also briefly surveyed CT and US and outlined their applications and limitations. For the case of MRI, the fundamental limitation is the acquisition speed. The section surveyed the advanced reconstruction techniques that can reduce the acquisition times. However, we also highlighted that these reconstruction approaches still require difficult hyper-parameter tuning to get the best image quality. In addition, as these approaches do not exploit available data or past scans fully, it leaves a room for improvement. In the next chapter, deep learning techniques are surveyed.

Chapter 3

Deep learning for medical imaging

3.1 Introduction

Deep learning [LBH15] is a sub-field of machine learning, which mainly studies *artificial neural network*. Deep learning gained its popularity in 2012 when it achieved the state-of-the-art performance for ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a large-scale image classification task including nearly 1000 different target classes [KSH12]. The basic theoretical foundation of artificial neural network has been laid out much earlier in 1960-90's [Ros61; R+88; L+95; HS97], however, the modern success of deep learning rooted in the availability of large scale data, improved techniques for building and extracting complex representation of data and the availability of the improved computational resources. This chapter provides the background on deep learning techniques and define the necessary terminologies that appear in this thesis. The deep learning techniques introduced in this thesis, in particular, *convolutional neural network* and *recurrent neural network* form the basis of the methods presented in the subsequent four chapters.

In general, there are three types of learning problems: supervised learning, unsupervised learning and reinforcement learning. In this thesis, we are mainly concerned with supervised learning approaches. The presentation is inspired by the statistical learning theory in [Vap13].

3.2 Supervised learning

The goal of supervised learning is to learn a mapping from an input space to an output space given training data of known input-output pairs. For example, for a given patient, we would like to predict whether the patient will develop congenital heart diseases (CHDs) in the next 5 years. In this case, the input variables may contain a set of patient information such as age, sex, height, weight, exercise frequency, etc., whereas the output can be a binary answer (yes or no), a probability (from 0 to 1) or the names of diseases. Supervised learning techniques allow one to train a model which attempts to perform such predictions from historical data of patients who did and did not develop CHDs.

We define the terminologies for this section. We denote an *input space* as $\mathcal{X}$, where we prescribe a *random variable* (r.v.) X and a *probability distribution* P_X to it.[1] Similarly, define an *output space* $\mathcal{Y}$, the corresponding r.v. Y and the probability distribution P_Y. The joint probability distribution on $\mathcal{X}\times\mathcal{Y}$ is defined as $P_{XY}(X,Y)$, which, by Bayes' rules, can be written a product of a *marginal* distribution $P_X(X)$ and a *conditional* distribution $P_{Y|X}(Y|X)$, i.e. $P_X(X)P_{Y|X}(Y|X)$.[2] Training data $\mathcal{D}$ is a finite set of input-output pairs: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \subseteq (\mathcal{X}\times\mathcal{Y})^N$, which are the samples from the joint distribution $P(X,Y)$. A set of possible mapping $h : \mathcal{X} \to \mathcal{Y}$ is called *hypothesis space* $\mathcal{H}$. The discrepancy between the true response y and the prediction $h(x)$ is quantified by a *loss* function $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_{\geq 0}$. An example of loss is Euclidean distance $\ell(y,y') = \|y - y'\|_2$.

The goal of supervised learning is to learn a mapping $h \in \mathcal{H}$, which takes $x \in \mathcal{X}$ as input and predict the most likely output $y \in \mathcal{Y}$, i.e. $h(x) = \text{argmax}_{y^*} P(Y = y^*|X = x) \approx y$. In a more mathematical terminology, the goal of supervised framework is to find a hypothesis (a mapping) $h \in \mathcal{H}$ which minimises *risk*:

$$R(h) = \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \ell(y, h(x)) dP(x,y), \tag{3.1}$$

[1] For completeness: we define a *probability space* $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mathbb{P}_\mathcal{X})$, where $\mathcal{X}$ is a *sample space*, $\mathcal{B}_\mathcal{X}$, called *event space*, is a Borel-algebra of $\mathcal{X}$, $\mathbb{P}_\mathcal{X} : \mathcal{B}_\mathcal{X} \to \mathbb{R}$ is the *probability measure* defined on event space, $X : \mathcal{X} \to \mathbb{R}$ is a random variable corresponding to $\mathbb{P}_\mathcal{X}$, i.e. $P_X : \mathbb{R} \to \mathbb{R}$ is a *push-forward measure*: $P_X(S \subseteq \mathbb{R}) = \mathbb{P}_\mathcal{X}(X^{-1}(S)) = \mathbb{P}_\mathcal{X}(\{\omega \in \mathcal{X} \text{ s.t. } X(\omega) \in S\})$.

[2] From here on, we drop the subscript on P if it is obvious: e.g. $P_Y(Y) = P(Y)$

i.e., the sum of loss over all input-output pairs weighted by the joint distribution. In practice, one does not have the access to the joint distribution $P(X, Y)$ but only limited numbers of training data $\mathcal{D}$, which captures the likely input-output pairs. In *empirical risk minimisation* (ERM) principle, the following functional is minimised instead:

$$R_{\text{emp}}(h) = \frac{1}{N} \sum_{i=1}^{N} \ell(y, h(x)) \tag{3.2}$$

In ERM principle, the model h is said to have *generalisation* property if minimising the empirical risk R_{emp} results in minimising the risk R. Generalisation is the key ingredient for making the model work on unseen cases.

The central theme of supervised learning is to find a suitable hypothesis *space* $\mathcal{H}$ (e.g. possible model configurations) and a *learning algorithm* $\mathcal{A}$, which is a sequence of steps that allows one to obtain a model that can generalise to unseen data points, all in a reasonable amount of computation. Currently there is no learning algorithms that can achieve this goal in a practical (computationally efficient) way, which makes the field of supervised learning largely based on empirical evidences. A field that aims to characterise and quantify the bounds for generalisation property is called statistical learning theory [Vap13] and it is still an ongoing area of research.

From a practical point of view, there are a few important concepts to have in mind when designing the hypothesis space $\mathcal{H}$ and the learning algorithm $\mathcal{A}$. The first concept is the *expressiveness* of the hypothesis space. Expressiveness describes whether the selected hypothesis space can capture a diverse range of the representations of data. For example, it may be possible to predict the likelihood of a patient developing CHDs using a linear model (e.g. a linear combination of input). However, for a large scale image classification task, linear models may not be able to capture the underlying nonlinear correlation. In this case, the expressiveness of linear models is limited. Secondly, a model is said to have a *robustness* property if the output does not change under a small perturbation of input and output. Another notion is *stability*, which is concerned with how consistently the model can be obtained from a perturbation in training data.

Finally, another important notion is *inductive bias*. Inductive bias is an assumption of the existence or the superiority of a certain class of algorithms over the other for a task. An illustrative example is *Occam's Razor*, which roughly states that, simple explanations should be preferred over the complex ones. Another example is, a certain class of $\mathcal{H}$ or a learning algorithm $\mathcal{A}$ may bias the model h to converge towards a minima of R_{emp} which generalises better. As such, the problem of generalisation is often rephrased as the problem of finding the correct inductive bias. It is believed that the techniques used in current deep learning have the relevant inductive bias that makes themselves generalisable for many tasks in visual, audio and language processing.

3.3 Deep learning and neural network

Deep Learning is a field where one uses *deep feed-forward neural networks* (DNN's) for the hypothesis space $\mathcal{H}$. Specifically, a feed-forward neural network is a parametric function $f : \mathcal{X} \times \Theta \to \mathcal{Y}$ given the parameter space Θ and is defined as:

$$f(x;\theta) = f_L(f_{L-1}(...f_1(f_0(x;\theta_0);\theta_1)...;\theta_{L-1});\theta_L) \tag{3.3}$$

$$f_i(\xi;\theta_i) = \sigma_i(W_i^T\xi + b_i), \quad \theta_i = \{W_i, b_i\}. \tag{3.4}$$

A neural network has learnable model parameters $\theta = \{W_0, b_0, W_1, b_1, \ldots, W_L, b_L\} \in \Theta$, which are called *weights*. Each f_i is called *layer* i. Typically, the input to each layer is expressed as a vector $x \in \mathbb{R}^{l_i}$. As such, W_i's and b_i's are 2D matrices and 1D vectors respectively. The dimensionality l_i of each layer (also called the number of features) can be arbitrarily specified by the designers. Note that if $f_i : \mathbb{R}^3 \to \mathbb{R}^{64}$ (a layer with $l_i = 64$ features), then the dimensions of the associated weights are $W_i \in \mathbb{R}^{3\times 64}$ and $b_i \in \mathbb{R}^{64}$. In addition to the network parameters, DNN's have many *hyper-parameters*, such as the depth L of the network, the number of features l_i, and the choice of *nonlinear activation functions* σ_i's. In general, f_i is called a fully-connected (FC) or dense layer if the 2D weight matrix has dense nonzero-entries. The evaluation of f on x is often called *forward-propagation*. Given empirical risk R_{emp}, the network is typically

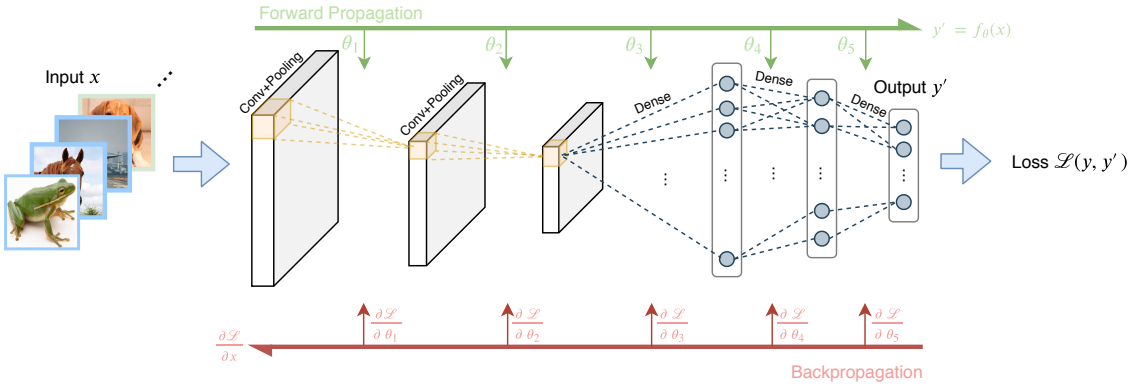

Figure 3.1: The schematic of how the network trained via backpropagation. The network architecture shown here is a convolutional neural network (see Section 3.4). Image adopted from [Sch+19c].

trained using gradient descent algorithm:

$$\theta = \theta - \alpha \nabla_\theta R_{\text{emp}} \tag{3.5}$$

where $\alpha \in \mathbb{R}$ is called a *step size* or a *learning rate*. The core idea of gradient descent is to iteratively find a small perturbation of the network parameters that can move the network performance towards lower empirical risk. Note that since the network has a multi-layer structure, the gradient can be computed in a layer-wise fashion via *chain rule*. For example, the gradient of the weight θ_i for the layer i can be computed as:

$$\nabla_{\theta_i} R_{\text{emp}} = \frac{\partial R_{\text{emp}}}{\partial f_L}^T \frac{\partial f_L}{\partial f_{L-1}}^T \cdots \frac{\partial f_i}{\partial \theta_i} \tag{3.6}$$

This update rule is often called *backpropagation*. The Jacobian of each layer $\frac{\partial f_L}{\partial f_{L-1}}^T$ can be computed by chain rule. Let f_i is the output of layer i and $z_i = W_i f_{i-1} + b_i$, then

$$\frac{\partial f_i}{\partial f_{i-1}}^T = \frac{\partial \sigma(z_i)}{\partial z_i}^T \frac{\partial (W_i f_{i-1} + b_i)}{\partial f_{i-1}}^T \tag{3.7}$$

Gradient descent requires evaluating the loss on all training dataset, which can be computationally expensive for large-scale problems. This is mitigated by performing *mini-batch* update, which computes R_{emp} using m data points instead, much smaller than the size of training data. The idea is that the risk computed with batch size m well-approximates the original empirical

risk R_{emp}. Secondly, α is often empirically chosen to be a small fixed number. This technique is called *stochastic gradient descent* (SGD).

DNN's are also expressive and it can represent a wide range of functions. To this end, we state *Kolomogorov-Arnold representation theorem* [BG09]:

Theorem 3.1. *Let $f : [0,1]^n \to \mathbb{R}$ be an arbitrary multivariate continuous function. Then it has the representation*

$$f(x_1, \ldots, x_n) = \sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^{n} \psi_{q,p}(x_p) \right) \tag{3.8}$$

with continuous one-dimensional inner and outer function Φ_q and $\psi_{q,p}$. All these functions Φ_q, $\psi_{q,p}$ are defined on the real line. The inner functions $\psi_{q,p}$ are independent of f.

Informally, the theorem states that a two-layer neural network is sufficient to express most functions, provided it has enough dimensionality. It is important to note that it does not provide a way of constructing such optimal neural network, neither guarantees that such network can be trained using SGD or other optimisation techniques. Therefore, while KAR theorem guarantees the expressiveness of the DNN's in theory, more heuristics are needed to make them work well in practice. As it turns out, convolutional neural network and recurrent neural network are DNN's which can implicitly provide a structural information, makes them effective for many tasks in computer vision, audio and natural language processing. These two techniques are now reviewed.

3.4 Convolutional neural network

Convolutional neural network (CNN) is a neural network architecture that has been shown to be extremely effective for numerous applications in computer vision in the recent years. Unlike the standard DNN's with FC layers, CNN first extracts features locally, which can build complex image representations while maintaining spatial correspondences. The building blocks

of CNN's are convolutional layers, nonlinearity layers, pooling layers, normalisation layers and (possibly) fully connected layers.

Convolutional layer A convolutional layer is a variant of FC layer where W has a form of *Toeplitz matrix*, i.e. the layer multiplication can be performed as a convolution:

$$W^T x = w \circledast x \tag{3.9}$$

More specifically, for an input tensor $\mathbf{x} \in \mathbb{R}^{N_x \times N_y \times N_c}$ (e.g. a 2D image with N_c channels) and a weight $\mathbf{w} \in \mathbb{R}^{k_x \times k_y \times N_c \times N_{c'}}$ and bias $\mathbf{b} \in \mathbb{R}^{N_{c'}}$, convolution layer f is defined as:

$$[f(\mathbf{x}; \mathbf{w}, \mathbf{b})]_{lmn} = \sum_{i=l-k_x/2}^{l+k_x/2} \sum_{j=m-k_y/2}^{m+k_y/2} \sum_{k=0}^{N_{c'}} \mathbf{w}_{ijkn} \mathbf{x}_{ijk} + \mathbf{b}_n \tag{3.10}$$

Weights in convolution layers are often called *convolutional kernels*. Convolutional layers have hyper-parameters such kernel width (k_x, k_y), number of features $N_{c'}$, *stride* and *dilation* factors. Compare to FC-layer in feed-forward networks, convolutional layer has much less parameters as the weight is shared across the image spatially. This enables the network to extract local features and build complex representations before they are aggregated for further analyses.

Nonlinearity layer The second component is a nonlinearity layer. A standard choice of a nonlinearity is *rectified-linear unit* (ReLU), which is defined element-wise as:

$$\mathrm{ReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

ReLU is a nonlinear activation function which solved *gradient vanishing problem* of other nonlinearities like sigmoid and tanh functions, enabling training of extremely deep networks. There are other variations such as leaky ReLU.

Pooling layer A *pooling layer* is used to locally aggregate the statistics of the intermediate features within a certain pooling window. If a pooling window size is $p_x \times p_y$, a *max-pooling* layer on $\mathbf{x}$ is defined as:

$$[\text{maxpool}(\mathbf{x})]_{lmc} = \max_{\substack{i\in[l-p_x/2,\ldots,l+p_x/2] \\ j\in[m-p_y/2,\ldots,m+p_y/2]}} \mathbf{x}_{ijc} \tag{3.12}$$

Since the layer aggregates the local values, CNN's with pooling layers are less sensitive to local perturbation. This attribute is often informally known as *translational invariance*. The second use case of pooling layer is to reduce the spatial dimension of input tensor. Often, the stride of the operation is matched by the pooling window. This meant that the output of pooling is down-scaled by a factor proportional to the pooling window size. Once the representation is down-scaled, the network can learn a representation which aggregates the local information and increasingly build a complex hierarchical representation. Besides from max-pooling, *average-pooling* is also used commonly.

Normalisation layer The final component is a normalisation layer. The most commonly used normalisation technique is *batch-normalisation* (BN) [IS15], which normalises the intermediate tensor by mini-batch statistics and have two learnable parameers γ, β. It first computes mean and variance of the batch, then applies them to whiten the feature representations. The learnable parameters are used to rescale the data:

$$\begin{aligned} \mu_{\mathcal{B}} &= \frac{1}{m}\sum_{i=1}^{m} x_i \\ \sigma^2_{\mathcal{B}} &= \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \\ \text{BN}_{\gamma,\beta}(x_i) &= \gamma\left(\frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + e}}\right) + \beta \end{aligned} \tag{3.13}$$

where m here is the batch size, x_i is i-th data in the batch. BN can help accelerate learning by normalising the statistics of each layer every time. The precise reasons of the effectiveness of BN is still not fully understood, although empirically it is observed that BN's allow larger learning

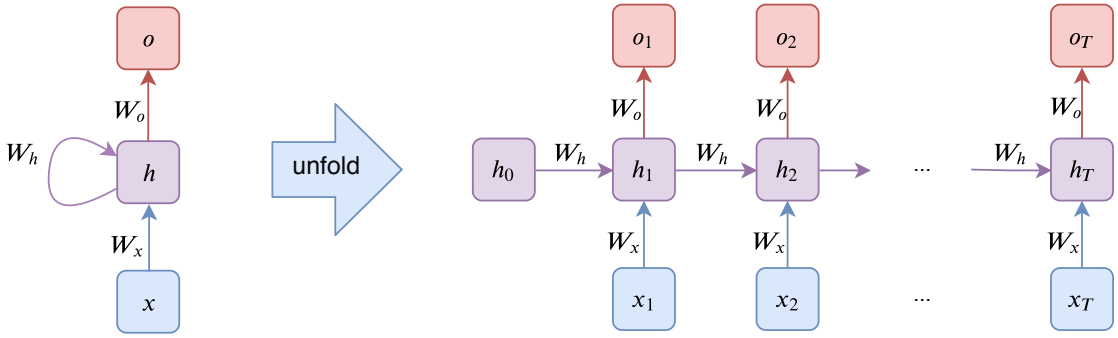

Figure 3.2: The schematic of an RNN architecture. Each arrow represents a multiplication with the weights and is followed by a nonlinearity layer which is not shown in the diagram. For RNN, the parameters are shared across the unfolding.

rate as well as provide robustness to some domain shift between train and test data [Bjo+18; San+18]. Besides from BN, other normalisation schemes exist, such as batch-renormalisation [Iof17], group normalisation, instance normalisation [UVL16] and layer normalisation [WH18].

3.5 Recurrent neural network

Recurrent neural network is designed such that features with long-term temporal dependency can be extracted. A vanilla RNN is usually expressed as following. Let x be our data. Suppose we have a temporal component to x, e.g. $x = (x_1, x_2, \ldots, x_T)$ (note that here the subscripts do not represent the indices in a batch, but the temporal indices). Then, suppose one desires to make a prediction at every time step t, then we have an RNN of the form:

$$\begin{aligned} h_0 &= 0 \\ h_t &= \sigma_h \left(W_x^T x_t + W_h^T h_{t-1} + b_h \right) \\ o_t &= \sigma_o \left(W_o^T h_t + b_o \right) \end{aligned} \tag{3.14}$$

where h_t is called *hidden state*, o_t is the output at time step t, $\theta = \{W_x, W_h, W_o, b_h, b_o\}$. In particular, RNNs combines the previous *hidden state* h_{t-1} and the current input x_t, to create a

new hidden state h_t. This structure enables the network to extract features with high temporal dependencies. RNN's are often used in natural language processing such as translation or sentence generation. For many tasks, o_i is not required and only the final output o_T is needed, such as text classification. A popular architecture includes bidiretional RNN [HWW15], gated recurrent units (GRUs) [Chu+14], *long-short term memory (LSTM)* [HS97].

3.6 Learning the network

CNN and RNN's are expressive and exhibits useful inductive biases for learning image features and temporally dependant features respectively. However, while these networks are extremely powerful, proper training methodologies are required to obtain good performances from them. In fact, the second aspect of the modern success of deep learning is the improve techniques to train these networks. This section lists the common tricks used to improve the network performance.

Optimiser Currently, most of the successful learning algorithms for large-scale optimisation problems (e.g. deep learning) are based on first-order *stochastic optimisation* techniques. As aforementioned, in SGD, a fixed learning rate and the minibatch update are two sources of stochasticity. The current understanding is that the stochasticity is important for the generalisation, as the stochastic nature of the training enables the networks to skip spurious minimums and allow the models to converge to a good local minimum. The exact reason for the effectiveness of stochastic first-order gradient optimisation method is still an active area of research.

There are many methods that have been proposed which try to improve upon either the convergence rate or generalisation property of the vanilla SGD. Essentially, the most important consideration is how to set the learning rate α at each iteration of gradient descent, often called *annealing schedule*. The problem is that the learning rate needs to be adaptive to avoid being stuck in slow update process as well as converging to poor local minimum, or if learning rate is too high, the method will diverge. Classically, methods such as SGD with momentum [Sut+13]

and Nesterov momentum [Nes83] have been popular for accelerating the convergence rate. However, for large scale non-convex optimisation problems such as training deep networks, the improvement from the momentum is not universally observed. More recently, advanced learning algorithms have been proposed which can adaptively adjust the learning rate. *RMSProp* is an optimiser [HSS12] which scales the gradient based on the current root-mean-square error. *Adam* [KB15] (predecessors are *AdaDelta* [Zei12] and *AdaMax* [DHS11]), is an optimiser which monitors the statistics (variance) to adjust the learning rate and momentum. While SGD with well-chosen annealing schedule works well in practice, the latter methods with adaptive learning rates are gaining popularity [Rud16]. Note that while the latter methods can adaptively change the learning rate, it is nevertheless often combined with annealing schedules.

Training, validating and testing Given dataset $\mathcal{D}$ with a suitable choice of architecture f, loss function ℓ and optimisation algorithm $\mathcal{A}$, one trains a network. In a typical setting, the available dataset is split into training, validation and testing subsets. The network is trained on training set, whereas small number of cases are reserved for validation set, often called held-out set. The aim of validation set is ensure that the network is not *overfitting* the training data (testing data is assumed inaccessible during training).

Usually one pass of gradient update is called *update iteration* and one pass through the entire dataset is called *epoch*. Usually, the training process is monitored by plotting R_{emp} against epoch. The *convergence criteria* refers to the mechanism to tell when the network training is done. *Early-stopping* is a technique where one terminates the training if validation error stops decreasing (i.e. to prevent overfitting).

Typically, the methods are evaluated using n-fold *cross-validation* process. In n-fold cross validation, the dataset is split into n small chunks, of which $n-1$ is used for training and the remaining is used for validation. The idea is that we don't draw conclusion after successful performance from particular validation data. By averaging out the result of cross-fold validation, we can get robust statistics.

Preventing overfitting Despite better models, it is well possible that the network overfit the training data. The techniques to avoid overfitting are called *regularisation* techniques. One popular technique is *weight regularisation* [GBC16]. Typically, this is added to training objective to constrain the magnitude of weights. The intuition is that highly nonlinear function is not robust to small perturbation, if its the weights have extremely large coefficients. Usually this constraints the norm of the weights θ by various norm such as ℓ_1, ℓ_2 norms. Note that adding these regularisation is equivalent to applying a prior to the weights $p(\theta)$, where for ℓ_2 norm, we have a prior that has zero-mean Gaussian with small variance. For ℓ_1 norm, this is equivalent to Laplace prior. Besides from regularising the norm, one also sometimes explicitly constraint the scale of norm, which are often called *clipping*.

Another very effective way of regularisation is *data augmentation*. The idea of data augmentation is to increase variation in training data so the network does not overfit on specific characteristic of the training data distribution, but also capture the plausible variation.

3.7 Deep learning for medical imaging

Due to its empirical success, deep learning methods are already widely applied for various medical imaging problems. While it is out-of-scope for the thesis to provide a comprehensive survey, this section highlights some of the work as illustrative examples of how deep learning is commonly applied for medical imaging problems. Numerous great surveys can be found at [Lit+17; RNZ18; LL19; Zah+18; Mai+19].

3.7.1 Medical image classification

Image classification is a process of predicting the label of a given input image. Formally this can be formulated as follows: each image $\mathbf{x} \in \mathcal{X}$ has an associated label $y \in \mathcal{Y}$, where y belongs to one of C target classes. To train a network for image classification tasks, one first expresses the labels as *one-hot encoding* vectors. A one-hot encoding vector $\mathbf{y}$ is a C-dimensional vector

such that if the label y belongs to i-th class, then $\mathbf{y}_i = 1$ and $\mathbf{y}_{j\neq i} = 0$. The task of the network is then to predict the one-hot encoding vector $\mathbf{y}$ given $\mathbf{x}$. In particular, the output of the network $\mathbf{y}_{\text{pred}} = q(y|\mathbf{x})$ is a C-dimensional probability vector, where the i-th entry expresses the likelihood of an input image $\mathbf{x}$ belonging to the i-th class. Since the output of the network is a probability vector, it sums to one. This behaviour is typically achieved through a *softmax* layer:

$$q(y = c|\mathbf{x}) = [\text{softmax}(\mathbf{z})]_c = \frac{e^{\mathbf{z}_c}}{\sum_{j=1}^{n} e^{\mathbf{z}_j}} \tag{3.15}$$

where $\mathbf{z}$ is the output of the network before softmax layer. Let $p(y|\mathbf{x})$ be the probability distribution of the image belonging to the target labels and Let $q(y|\mathbf{x})$ be the network output. Then, the network is trained using *cross-entropy* loss, which is defined as:

$$\ell_{\text{CE}}(p, q) = -\sum_{i=0}^{C} p(\mathbf{y} = i|\mathbf{x}) \log q(y = i|\mathbf{x}) \tag{3.16}$$

In most scenarios, each image belongs to one class (i.e. if $\mathbf{x}$ has label c, $p(y = c|\mathbf{x}) = 1$ and 0 otherwise), the loss simplifies to:

$$\ell_{\text{CE}}(p, q) = -\log q(y = c|\mathbf{x}) \tag{3.17}$$

Neural networks can then be trained using the strategy defined in Section 3.6. Image classification is a task that popularised the deep learning approaches when it achieved the state-of-the-art performance for ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [KSH12]. In medical image analysis, image classification has a wide range of important applications. This includes exam classification which can be used for screening [KCB16; Ant+16; Cio+17; WB16], lesion class or pathology classifications [Spa+16; GHT17; She+15; Est+17; Clo+19], image scan plane classification for ultrasound [Bau+16; Sch+18b] and MRI, as well as classifying image quality or image artefact [Oks+18; Oks+19].

Compared to computer vision problems, medical image classification problems have their unique challenges. Firstly, it is non-trivial to acquire a large number of data containing pathology,

so there is a need for addressing a class imbalance problem. To address this, often transfer learning is employed to train a deep network [KCB16; Ant+16]. Typically in transfer learning, the networks trained for large scale image classification tasks are then finetuned for the target application. The most standard choices of the architectures are VGG [SZ14], Resnet [He+15] and DenseNet [Hua+17]. The second challenge is that the medical images are often three dimensional or non-Euclidean, so often it is the case the networks are extended to three-dimensional [HGE16] or even on non-Euclidean graphs [Ars+18]. Finally, the object that contributes to the image label (e.g. lesion) is extremely small compared with respect to the entire image [ZDG19; Paw+19]. To address this, features at multiple levels are often aggregated to improve the performance [Dou+16]. In general, the trend in image classification follows the trend in computer vision. For the extensive list of applications, refer to [Lit+17].

3.7.2 Medical image segmentation

The goal of image segmentation is to classify each pixel in the input image into one of the C target classes. More formally, for an input image $\mathbf{x} \in \mathbb{R}^{N_x \times N_y}$, the goal is to obtain the segmentation map $\mathbf{y} \in \mathbb{R}^{N_x \times N_y \times C}$. Thus, the network outputs C-dimensional probability vector for each pixel: $\mathbf{y}_{\text{pred}-i} = q(y_i = c|\mathbf{x}_i)$ where y_i is the label for i-th pixel. The loss function that can be used for segmentation is similar to classification, except that the label prediction is now obtained per pixel and the loss is aggregated for all pixels:

$$\ell_{\text{CE}}(p, q) = -\sum_{i}^{N_x \times N_y} \log q(y_i = c|\mathbf{x}_i) \tag{3.18}$$

Image segmentation is one of the fundamental tasks in medical image analysis due to the range of application areas: organ and structure segmentation [Bai+18; SEG17a; Cer+18b; Sin+18] and legion segmentation [Kam+17; CBR17; Bow+17; Che+18]. The segmented structures serve as important features for further analyses [Bel+19]. For segmentation tasks, most architecture is based on multi-scale analyses. This include FCN [Bai+18], Unet [RFB15] and DeepMedic [Kam+17]. *Ensembling* of the networks is also an effective technique to get stable results [Kam+18]. In many cases, the deep learning based segmentation achieves human level

performance for many application areas, i.e., it is within the inter-observer range of the manual annotators [Bai+18; Sin+18; Ber+18].

The main challenges of medical image segmentation is similar to image classification. Firstly, there is a problem of class imbalance in the target labels [LKG19; Has+18]. For example, lesions are often much smaller than the background. As larger structures have more pixels associated with it, a network trained with cross entropy loss tends to bias the predictions towards getting the larger structures correctly. Dice loss [Sud+17] and their extensions [SEG17b] are considered to be more robust to small structures.

Secondly, collecting a large dataset with manual annotations is challenging, so often one needs to work with small dataset. A various approach has been proposed, including weakly supervised segmentation [Raj+16; Can+18], as well as data augmentation approaches [Che+19a; Cha+19b; Bow+18]. Another key problem in medical imaging is that even for images of the same underlying anatomy, the image statistics can differ based on the scanner type (MRI/CT), image contrast, noise and resolution and the networks can fail to generalise across unseen scanner protocols. To this end, many approaches to extract common feature has been proposed [Che+19c; Dou+18; Ouy+19a].

An interesting property of medical image analysis that is different from computer vision is that the target anatomy is often highly structured. Many approaches exploit *shape priors* [Bif+19; Cer+18a]. In [Okt+17], low-dimensional representation of the anatomy is first learnt to anatomically constrain the network reconstruction via loss function. In [Che+19b], 3D features of the anatomy is learnt, which is combined to improve the segmentation from multiple 2D views. In [Qin+18a; Qin+18b], the motion field constraint is used to improve CMR segmentation. The shape prior approaches are also combined with traditional optimisation approaches, such as atlas-based registration [Dua+19].

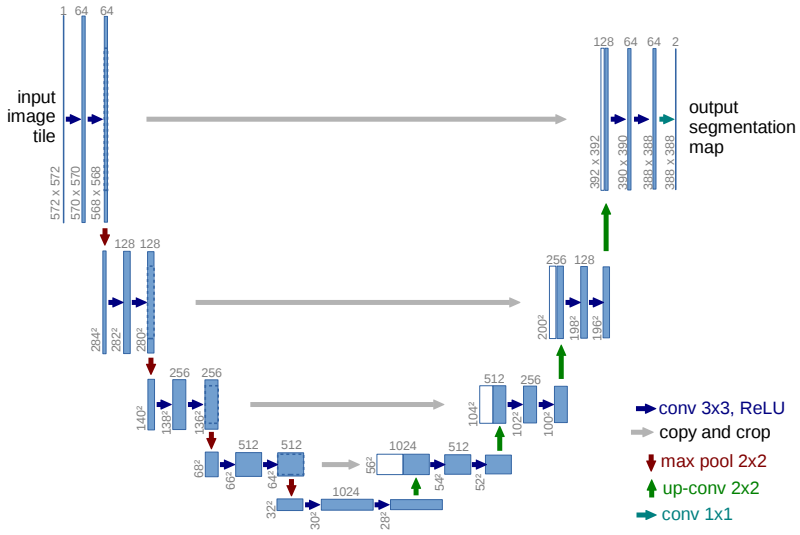

Figure 3.3: U-net architecture proposed by Ronneberger et al. for image segmentation [RFB15].

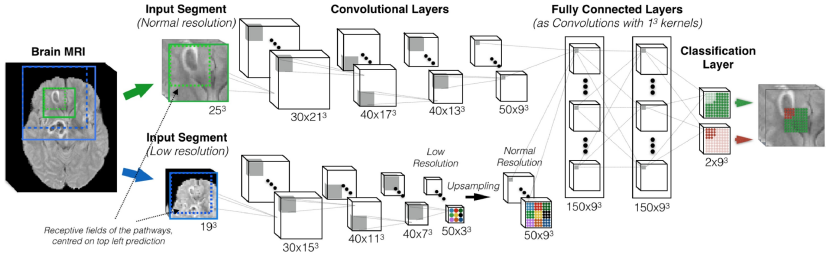

Figure 3.4: Deep Medic architecture proposed by Kamnitsas et al. for brain lesion segmentation [Kam+17].

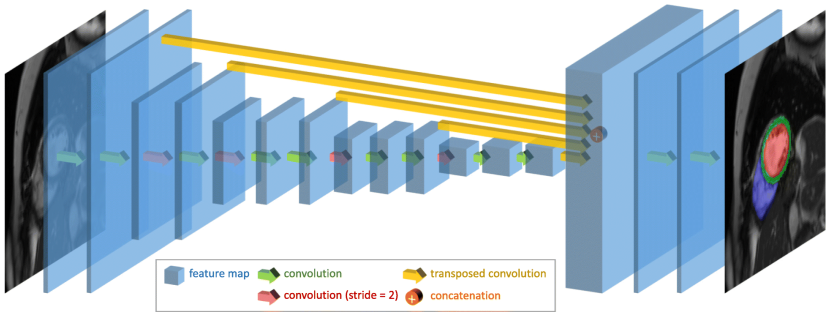

Figure 3.5: FCN architecture used by Bai et al. for cardiac image segmentation [Bai+18].

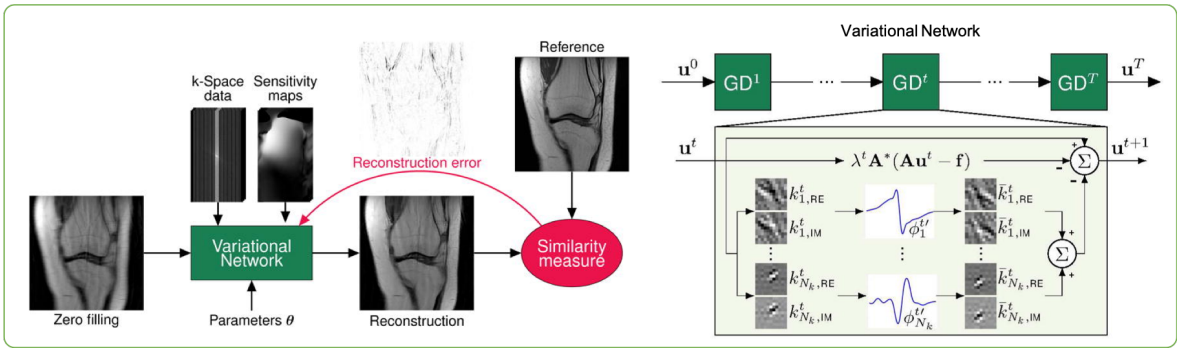

Figure 3.6: Variational network architecture proposed by Hammernik et al. for accelerated MR image reconstruction [Ham+18].

3.7.3 Medical image reconstruction

Image reconstruction is another branch were deep learning is making a significant impact. As briefly surveyed in Chapter 2, image reconstruction problems generally fall under the category of *inverse problem*, which includes wide class of problems such as image denoising, inpainting, deblurring, single-image super resolution, accelerated MR reconstruction, CT and PET reconstruction, and so on.

Accelerated MR image reconstruction In general, currently there are mainly three different types of approaches for deep learning based MR image reconstruction. The first approach is an end-to-end learning approach, which, to the best of our knowledge, was brought to attention by Wang et al. [Wan+17b]. Wang et al. demonstrated that the undersample images could be reconstructed using 3-layer CNN. In the following year, more powerful networks were employed, mainly based on U-net type architectures [Han+18]. At the same time, Zhu et al. proposed AUTOMAP [Zhu+18a] to learn image reconstruction directly from k-space using a combination of FC layers and convolution layers. More recently, Han et al. proposed to reconstruct undersampled MR images directly in k-space, followed by inverse Fourier transform [HY18]. The latter method is inspired by the connection between low-rank approaches and the deep learning architecture.

The second class of approaches is called an unrolled method. To the best of our knowledge, Hammernik et al. proposed a network architecture called variational network (VN) [Ham+18].

The idea of unrolled approach is to first formulate the general inverse problem as in traditional optimisation algorithm (c.f. Eq. (2.13)), then explicitly write out the optimisation algorithm, such as gradient descent. It turns out that under certain assumptions on the regularisation terms, the gradient descent steps can be explicitly written as a convolutional layers and nonlinearity layers. From this observation, a network architecture, where optimisation step is *unrolled*, is proposed. Different optimisation algorithms can be "unrolled", which results in different network architectures. For example, alternate direction methods of multipliers (ADMM) is extended to Deep-ADMM network [S+16]. Shortly after, we proposed *Deep Cascade of CNN's* (DC-CNN), which was inspired by unrolling of dictionary learning MRI approaches [RB11]. Later on, the model was theorised as an unrolling of *variable-splitting* approaches [Qin+19]. A number of variations have been proposed, such as a model based on proximal update rules [Mar+17] as well as extension of DC-CNN to parallel image reconstruction [AMJ17].[3]

Finally, the third type of approaches is to combine deep learning with traditional optimisation algorithms. In this case, the regularisation $\mathcal{R}$ in Eq. (2.13) is replaced by prior based on deep learning. [Zha+17a; Ric+17; TBK17]. Tezcan et al. have shown that the method is highly adaptive to unseen data corruption [TBK17]. Alternatively, deep neural network architecture can be directly used as a constraint in GRAPPA type approaches [Akc+18].

Another important consideration for the reconstruction task is the loss function for training the network. A standard choice is ℓ_2 loss between the target image and the reconstruction from the neural networks. However, for image restoration tasks, ℓ_2 is often associated with blurring. As such, ℓ_1 loss or structural similarity (SSIM) loss are preferred [Zha+16; Ham+17b]. SSIM is a reference-based image quality metric and for two images x and y, it is defined as:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)\,(2\sigma_{xy} + C_2)}{\left(\mu_x^2\mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{3.19}$$

where $\mu_{(\cdot)}$, $\sigma_{(\cdot)}$ are the local mean and variance of each pixel in the image respectively and

[3] While one of the main contributions of the thesis is the development of the novel deep learning architectures for MR image reconstruction, a considerable amount of progress has been made at the same time. As such, this section is written to be chronologically up-to-date, although comparison with all other methods within the thesis is out-of-scope. Nevertheless, most of the proposed approaches mentioned here only addresses 2D image reconstruction, whereas the subsequent chapters are concerned with dynamic MRI applications.

σ_{xy} is the local covariance of each pixel in x and y. Another notable approach for improving the sharpness of the image is to use generative adversarial networks (GAN) [Goo+14]. GAN is a framework that consists of two networks, *generator* and *discriminator*. The general idea behind GAN is the discriminator D_{θ_D} is designed to distinguish (classify) between real images $x \sim p_{\mathcal{X}}(x)$ and the network generated images $x_{\text{rec}} \sim p_G(x)$, whereas the generator G_{θ_G} is trained to fool the discriminator – i.e. the generator tries to compromise the classification accuracy of the discriminator D. In this way, the generator can learn to create images that appears like real images from $p_{\mathcal{X}}(x)$. The adversarial training can be done by minimising the following objective function:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{x \sim p_{\mathcal{X}}(x)} \left[\log D_{\theta_D}(x)\right] + \mathbb{E}_{x_{\text{rec}} \sim p_G(x)} \left[\log(1 - D_{\theta_D}(G_{\theta_G}(x_{\text{rec}})))\right] \tag{3.20}$$

For image restoration tasks, the input to the generator is corrupted image. In practice, GAN loss is often combined with perceptual loss [JAF16; Led+17]. Several authors have applied GAN-based approach for MR reconstruction [Yan+18; Sei+18; Mar+19a; QNJ18]. However, one issue associated with GAN is the difficulty of evaluating the effectiveness of it. While GAN can generate sharp, realistic looking reconstructions, it can hallucinate features if the network is over-parameterised [Yan+18; CLH18]. Therefore, while these models often outperform the networks without GAN in terms of mean opinion scores (MOS) of the radiologists, some effort is required to guarantee that hallucination does not occur. In particular, currently there is no rigorous theory which characterises the behaviour of GAN.

Most deep learning approaches are shown to outperform traditional compressed sensing based approaches. However, currently there is no general consensus on what the optimal architecture is. To this end, Zbontar et al. [Zbo+18] have organised a large-scale accelerated MR image reconstruction challenge, which enables the comparison of the submitted models.

Finally, we note that neural network based reconstruction approaches are also adopted for reduced dosage CT [Wur+18; Ham+17a; Jin+17; Hau+19] and PET reconstructions [Gon+18], including both the unrolled approaches as well as the direct reconstruction approaches.

3.7.4 Summary

This section surveyed the deep learning techniques which are the core of the thesis. We first provided generic definition of the supervised learning framework, then presented the theory of neural network. We then surveyed the techniques that led to the success of modern deep learning. We then highlighted some of the successful applications of deep learning techniques for medical image analysis, including classification, segmentation and reconstruction. The subsequent chapters will now present the proposed deep learning approaches which attempt to solve the current insufficiency in medical imaging pipelines.

Chapter 4

Convolutional neural network for dynamic MRI reconstruction

This section is based on the following publications:

- **Schlemper, J.**, Caballero, J., Hajnal, J. V., Price, A. N., Rueckert, D. (2017). A deep cascade of convolutional neural networks for MR image reconstruction. Abstract 0643, 25th Annual Meeting and Exhibition International Society of Magnetic Resonance in Medicine, 2017.

- **Schlemper, J.**, Caballero, J., Hajnal, J. V., Price, A., Rueckert, D. (2017, June). A deep cascade of convolutional neural networks for MR image reconstruction. In International Conference on Information Processing in Medical Imaging (pp. 647-658). Springer, Cham.

- **Schlemper, J.**, Caballero, J., Hajnal, J. V., Price, A. N., Rueckert, D. (2017). A deep cascade of convolutional neural networks for dynamic MR image reconstruction. IEEE transactions on Medical Imaging, 37(2), 491-503.

4.1 Introduction

In many clinical scenarios, medical imaging is an indispensable diagnostic and research tool. One such important modality is Magnetic Resonance Imaging (MRI), which is non-invasive and offers excellent resolution with various contrast mechanisms to reveal different properties of the underlying anatomy. However, MRI is associated with an inherently slow acquisition process. This is because data samples of an MR image are acquired sequentially in *k-space* and the speed at which k-space can be traversed is limited by physiological and hardware constraints [Lus+08]. A long data acquisition procedure imposes significant demands on patients, making this imaging modality expensive and less accessible. One possible approach to accelerate the acquisition process is to undersample k-space, which in theory provides an acceleration rate proportional to a reduction factor of a number of k-space traversals required. However, undersampling in k-space violates the Nyquist-Shannon theorem and generates aliasing artefacts when the image is reconstructed. The main challenge in this case is to find an algorithm that can recover an uncorrupted image taking into account the undersampling regime combined with a-priori knowledge of appropriate properties of the image to be reconstructed.

Using Compressed Sensing (CS), images can be reconstructed from sub-Nyquist sampling, assuming the following: firstly, the images must be *compressible*, i.e. they have a sparse representation in some transform domain. Secondly, one must ensure *incoherence* between the sampling and sparsity domains to guarantee that the reconstruction problem has a unique solution and that this solution is attainable. In practice, this can be achieved with random subsampling of k-space, which produces aliasing patterns in the image domain that can be regarded as correlated noise. Under such assumptions, images can then be reconstructed through nonlinear optimisation or iterative algorithms. The class of methods which apply CS to the MR reconstruction problem is termed CS-MRI [Lus+08]. In general, these methods use a fixed sparsifying transforms, e.g. wavelet transformations. A natural extension of these approaches has been to enable more flexible representations with *adaptive* sparse modelling, where one attempts to learn the optimal sparse representation from the data directly. This can be done by exploiting, for example, dictionary learning (DL) [RB11].

Figure 4.1: An example of the image acquisition with Cartesian undersampling for a sequence of cardiac cine images. (a) A ground truth sequence that is fully-sampled in k-space, shown along x-y and y-t for the image frame and the temporal profile respectively. (b) A Cartesian undersampling mask that only acquires 1/12 of samples in k-space, where white indicates the sampled lines. Each image frame is undersampled with the mask shown along k_x-k_y. The undersampling pattern along the temporal dimension is shown in k_y-t. (c) The zero-filled reconstruction of the image acquired using the 12-fold undersampling mask. (d, e) 4-fold Cartesian undersampling mask and the resulting zero-filled image. Note that the aliasing artefact becomes more prominent as the undersampling factor is increased.

To achieve more aggressive undersampling, several strategies can be considered. One way is to further exploit the inherent redundancy of the MR data. For example, in dynamic imaging, one can make use of spatio-temporal redundancies [Cab+14b], [JYK07], [QJ16], or when imaging a full 3D volume, one can exploit redundancy from adjacent slices [Hir+15]. An alternative approach is to exploit sources of explicit redundancy of the data to turn the initially underdetermined problem arising from undersampling into a determined or overdetermined problem that is easily solved. This is the fundamental assumption underlying parallel imaging [Uec+14]. Similarly, one can make use of multi-contrast information [HCA14] or the redundancy generated by multiple filter responses of the image [PL14]. These explicit redundancies can also be used to complement the sparse modelling of inherent redundancies [JLY15], [Lia+09].

Recently, deep learning has been successful at tackling many computer vision problems. Deep neural network architectures, in particular convolutional neural networks (CNNs), are becoming the state-of-the-art technique for various imaging problems including image classification [He+16], object localisation [Ren+15] and image segmentation [RFB15]. Deep architectures are capable of extracting features from data to build increasingly abstract representations, replacing the traditional approach of carefully hand-crafting features and algorithms. For example, it has already been demonstrated that CNNs outperform sparsity-based methods in super-resolution [Don+14] in terms of both reconstruction quality and speed [Shi+16]. One of the contributions

of our work is to explore the application of CNNs in undersampled MR reconstruction and investigate whether they can exploit data redundancy through learned representations. In fact, CNNs have already been applied to compressed sensing from random Gaussian measurements [Kul+16]. Despite the popularity of CNNs, there has only been preliminary research on CNN-based MR image reconstruction [Ham+18], [S+16], [Wan+16], hence the applicability of CNNs to this problem for various imaging protocols has yet to be fully explored.

In this work we consider reconstructing dynamic sequences of 2D cardiac MR images with Cartesian undersampling, as well as reconstructing each frame independently, using CNNs. We view the reconstruction problem as a de-aliasing problem in the image domain. Reconstruction of undersampled MR images is challenging because the images typically have low signal-to-noise ratio, yet often high-quality reconstructions are needed for clinical applications. To resolve this issue, we propose a deep network architecture which forms a cascade of CNNs.[1] Our cascade network closely resembles the iterative reconstruction of DL-based methods, however, our approach allows end-to-end optimisation of the reconstruction algorithm. For 2D reconstruction, the proposed method is compared to *Dictionary Learning MRI (DLMRI)* [RB11] and for dynamic reconstruction, the method is compared to Dictionary Learning with Temporal Gradient (*DLTG*) [Cab+14b], kt Sparse and Low-Rank (kt-SLR) [Lin+11] and *Low-Rank Plus Sparse Matrix Decomposition (L+S)* [OCS15], which are the state-of-the-art compressed sensing and low-rank approaches. We show that the proposed method outperforms them in terms of reconstruction error and perceptual quality, especially for aggressive undersampling rates. Moreover, owing to GPU-accelerated libraries, images can be reconstructed efficiently using the approach. In particular, for 2D reconstruction, each image can be reconstructed in about 23ms, which is fast enough to enable real-time applications. For the dynamic case, sequences can be reconstructed within 10s, which is reasonably fast for off-line reconstruction methods.

[1] Code available at `https://github.com/js3611/Deep-MRI-Reconstruction`

4.2 Problem formulation

Let $\mathbf{x} \in \mathbb{C}^N$ represent a sequence of 2D complex-valued MR images stacked as a column vector, where $N = N_x N_y N_t$. Our problem is to reconstruct $\mathbf{x}$ from $\mathbf{y} \in \mathbb{C}^M$ ($M \ll N$), undersampled measurements in k-space, such that:

$$\mathbf{y} = \mathbf{F}_u \mathbf{x} + \mathbf{e} \tag{4.1}$$

Here $\mathbf{F}_u \in \mathbb{C}^{M \times N}$ is an undersampled Fourier encoding matrix and $\mathbf{e} \in \mathbb{C}^M$ is acquisition noise modelled as additive white Gaussian (AWG) noise. In the case of Cartesian acquisition, we have $\mathbf{F}_u = \mathbf{M}\mathbf{F}$, where $\mathbf{F} \in \mathbb{C}^{N \times N}$ applies two-dimensional Discrete Fourier Transform (DFT) to each frame in the sequence and $\mathbf{M} \in \mathbb{C}^{M \times N}$ is an undersampling mask selecting lines in k-space to be sampled for each frame. The corresponding subset of indices sampled in k-space is indicated by Ω. For the fully-sampled case, $M = N$, the sequence is reconstructed by applying the 2D inverse DFT (IDFT) to each frame. However, Eq. (4.1) is underdetermined even in the absence of noise, and hence the inversion is ill-posed; in particular, applying IDFT, which in this case is also called *zero-filled* reconstruction, results in a sequence of aliased images $\mathbf{x}_u = \mathbf{F}_u^H \mathbf{y}$ due to sub-Nyquist sampling. Note that $\mathbf{F}_u^H$ is the Hermitian of the encoding matrix, which first maps $\mathbf{y} \in \mathbb{C}^M$ to the k-t coordinate and then applies the 2D IDFT frame-wise. Examples of the aliased images are shown in Fig. 4.1. Therefore, in order to reconstruct $\mathbf{x}$, one must exploit a-priori knowledge of its properties, which can be done by formulating an unconstrained optimisation problem:

$$\min_{\mathbf{x}} \quad \mathcal{R}(\mathbf{x}) + \lambda \|\mathbf{y} - \mathbf{F}_u \mathbf{x}\|_2^2 \tag{4.2}$$

$\mathcal{R}$ expresses regularisation terms on $\mathbf{x}$ and $\lambda \in \mathbb{R}$ allows the adjustment of data fidelity based on the noise level of the acquired measurements $\mathbf{y}$. For CS-based methods, the regularisation terms $\mathcal{R}$ typically involve ℓ_0 or ℓ_1 norms in the sparsifying domain of $\mathbf{x}$. Our formulation is inspired by DL-based reconstruction approaches [RB11], in which the problem is formulated as:

$$\min_{\mathbf{x},\mathbf{D},\{\boldsymbol{\gamma}_i\}} \quad \sum_i \left(\|\mathbf{R}_i \mathbf{x} - \mathbf{D}\boldsymbol{\gamma}_i\|_2^2 + \nu \|\boldsymbol{\gamma}_i\|_0 \right) + \lambda \|\mathbf{y} - \mathbf{F}_u \mathbf{x}\|_2^2 \tag{4.3}$$

Here $\mathbf{R}_i$ is an operator which extracts a spatio-temporal image patch at i, $\boldsymbol{\gamma}_i$ is the corresponding sparse code with respect to a dictionary $\mathbf{D}$. In this approach, the regularisation terms force $\mathbf{x}$ to be approximated by the reconstructions from the sparse code of patches. By taking the same approach, for our CNN formulation, we force $\mathbf{x}$ to be well-approximated by the CNN reconstruction:

$$\min_{\mathbf{x}} \quad \|\mathbf{x} - f_{\text{cnn}}(\mathbf{x}_u|\boldsymbol{\theta})\|_2^2 + \lambda\|\mathbf{F}_u\mathbf{x} - \mathbf{y}\|_2^2 \tag{4.4}$$

Here f_{cnn} is the forward mapping of the CNN parameterised by $\boldsymbol{\theta}$, possibly containing millions of adjustable network weights, which takes in the zero-filled reconstruction $\mathbf{x}_u$ and directly produces a reconstruction as an output. Since $\mathbf{x}_u$ is heavily affected by aliasing from sub-Nyquist sampling, the CNN reconstruction can therefore be seen as solving a de-aliasing problem in the image domain. The approach of Eq. (4.4), however, is limited in the sense that the CNN reconstruction and the data fidelity are two independent terms. In particular, since the CNN operates in the image domain, it is trained to reconstruct the sequence without a-priori information of the acquired data in k-space. However, if we already know some of the k-space values, then the CNN should be discouraged from modifying them, up to the level of acquisition noise. Therefore, by incorporating the data fidelity in the learning stage, the CNN should be able to achieve better reconstruction. This means that the output of the CNN is now conditioned on Ω and λ. Then, our final reconstruction is given simply by the output, $\mathbf{x}_{\text{cnn}} = f_{\text{cnn}}(\mathbf{x}_u|\boldsymbol{\theta}, \lambda, \Omega)$. Given training data $\mathcal{D}$ of input-target pairs $(\mathbf{x}_u, \mathbf{x}_{\text{gnd}})$ where $\mathbf{x}_{\text{gnd}}$ is a fully-sampled ground-truth data, we can train the CNN to produce an output that attempts to accurately reconstruct the data by minimising an objective function:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{(\mathbf{x}_u, \mathbf{x}_{\text{gnd}}) \in \mathcal{D}} \ell\left(\mathbf{x}_{\text{gnd}}, \mathbf{x}_{\text{cnn}}\right) \tag{4.5}$$

where ℓ is a loss function. In this work, we consider an element-wise squared loss, which is given by $\ell\left(\mathbf{x}_{\text{gnd}}, \mathbf{x}_{\text{cnn}}\right) = \|\mathbf{x}_{\text{gnd}} - \mathbf{x}_{\text{cnn}}\|_2^2$.

4.3 Data consistency layer

Denote the Fourier encoding of the image reconstructed by CNN as $\mathbf{s}_{\text{cnn}} = \mathbf{F}\mathbf{x}_{\text{cnn}} = \mathbf{F}f_{\text{cnn}}(\mathbf{x}_u|\boldsymbol{\theta})$. $\mathbf{s}_{\text{cnn}}(j)$ represents an entry at index j in k-space. The undersampled data $\mathbf{y} \in \mathbb{C}^M$ can be mapped onto the vectorised representation of k-t coordinate ($\mathbb{C}^N$) by $\mathbf{s}_0 = \mathbf{F}\mathbf{F}_u^H\mathbf{y}$, which fills the non-acquired indices in k-space with zeros. In order to incorporate the data fidelity in the network architecture, we first note the following: for fixed network parameters $\boldsymbol{\theta}$, Eq. (4.4) has a closed-form solution $\mathbf{s}_{\text{rec}}$ in k-space, given as in [RB11] element-wise:

$$\mathbf{s}_{\text{rec}}(j) = \begin{cases} \mathbf{s}_{\text{cnn}}(j) & \text{if } j \notin \Omega \\ \frac{\mathbf{s}_{\text{cnn}}(j)+\lambda\mathbf{s}_0(j)}{1+\lambda} & \text{if } j \in \Omega \end{cases} \tag{4.6}$$

The final reconstruction in the image domain is then obtained by applying the inverse Fourier encoding $\mathbf{x}_{\text{rec}} = \mathbf{F}^H\mathbf{s}_{\text{rec}}$. The solution yields a simple interpretation: if the k-space coefficient $\mathbf{s}_{\text{rec}}(j)$ is initially unknown (i.e. $j \notin \Omega$), then we use the predicted value from the CNN. For the entries that have already been sampled ($j \in \Omega$), we take a linear combination between the CNN prediction and the original measurement, weighted by the level of noise present in $\mathbf{s}_0$. In the limit $\lambda \to \infty$ we simply replace the j-th predicted coefficient in Ω by the original coefficient. For this reason, this operation is called a *data consistency step* in k-space (DC). In the case of where there is non-neglegible noise present in the acquisition, $\lambda = q/\sigma$ must be adjusted accordingly, where q is a hyper-parameter and σ^2 is the power of AWG noise in k-space (i.e. $\Re(\mathbf{e}_i), \Im(\mathbf{e}_i) \sim N(0, \sigma/\sqrt{2})$). In [Cab+14b], it is empirically shown that $p \in [5 \times 10^{-5}, 5 \times 10^{-6}]$ for $\sigma^2 \in [4 \times 10^{-8}, 10^{-9}]$ works sufficiently well.

Since the DC step has a simple expression, we can in fact treat it as a layer operation of the network, which we denote as a *DC layer*. When defining a layer of a network, the rules for forward and backward passes must be specified in order for the network to be end-to-end trainable. This is because CNN training can effectively be performed through stochastic gradient descent, where one updates the network parameters $\boldsymbol{\theta}$ to minimise the objective function $\mathcal{L}$ by descending along the direction given by the derivative $\partial\mathcal{L}/\partial\boldsymbol{\theta}^T$. Therefore, it is necessary to

define the gradients of each network layer relative to the network's output. In practice, one uses an efficient algorithm called *backpropagation* [R+88], where the final gradient is given by the product of all the Jacobians of the layers contributing to the output. Hence, in general, it suffices to specify a layer operation f_L for the forward pass and derive the Jacobian of the layer with respect to the layer input $\partial f_L / \partial \mathbf{x}^T$ for the backward pass.

Forward pass The data consistency in k-space can be simply decomposed into three operations: Fourier transform $\mathbf{F}$, data consistency f_{dc} and inverse Fourier transform $\mathbf{F}^H$. The data consistency f_{dc} performs the element-wise operation defined in Eq. (4.6), which, assuming $\mathbf{s}_0(j) = 0 \; \forall j \notin \Omega$, can be written in matrix form as:

$$f_{dc}(\mathbf{s}, \mathbf{s}_0; \lambda) = \mathbf{\Lambda}\mathbf{s} + \frac{\lambda}{1+\lambda}\mathbf{s}_0 \tag{4.7}$$

Here $\mathbf{\Lambda}$ is a diagonal matrix of the form:

$$\mathbf{\Lambda}_{kk} = \begin{cases} 1 & \text{if } j \notin \Omega \\ \frac{1}{1+\lambda} & \text{if } j \in \Omega \end{cases} \tag{4.8}$$

Combining the three operations defined above, we can obtain the forward pass of the layer performing data consistency in k-space:

$$f_L(\mathbf{x}, \mathbf{y}; \lambda) = \mathbf{F}^H \mathbf{\Lambda} \mathbf{F} \mathbf{x} + \frac{\lambda}{1+\lambda}\mathbf{F}_u^H \mathbf{y} \tag{4.9}$$

Backward pass In general, one requires *Wirtinger calculus* to derive a gradient in complex domain [Ami+11]. However, in our case, the derivation greatly simplifies due to the linearity of the DFT matrix and the data consistency operation. The Jacobian of the DC layer with respect to the layer input $\mathbf{x}$ is therefore given by:

$$\frac{\partial f_L}{\partial \mathbf{x}^T} = \mathbf{F}^H \mathbf{\Lambda} \mathbf{F} \tag{4.10}$$

Note that unlike many other applications where CNNs process real-valued data, MR images are complex-valued and the network needs to account for this. One possibility would be to design the network to perform complex-valued operations. A simpler approach, however, is to accommodate the complex nature of the data with real-valued operations in a dimensional space twice as large (i.e. we replace $\mathbb{C}^N$ by $\mathbb{R}^{2N}$). In the latter case, the derivations above still hold due to the fundamental assumption in Wirtinger calculus.

The DC layer has one hyperparameter $\lambda \in \mathbb{R}$. This value can be fixed or made trainable. In the latter case, the derivative $\frac{\partial f_{\mathrm{dc}}}{\partial \lambda}$ (a column vector here) is given by:

$$\left[\frac{\partial f_{\mathrm{dc}}(\mathbf{s}, \mathbf{s}_0; \lambda)}{\partial \lambda}\right]_j = \begin{cases} 0 & \text{if } j \notin \Omega \\ \frac{\mathbf{s}_0(j) - \mathbf{s}_{\mathrm{cnn}}(j)}{(1+\lambda)^2} & \text{if } j \in \Omega \end{cases} \tag{4.11}$$

and the update is $\Delta\lambda = \mathbf{J}_e \frac{\partial f_{\mathrm{dc}}}{\partial \lambda}$ where $\mathbf{J}_e$ is the error backpropagated via the Jacobians of the layers proceeding f_{dc}.

4.4 Cascading network

For CS-based methods, in particular for DL-based methods, the optimisation problem such as in Eq. (4.3) is solved using a coordinate-descent type algorithm, alternating between the de-aliasing step and the data consistency step until convergence. In contrast, with CNNs, we are performing one step de-aliasing and the same network cannot be used to de-alias iteratively. While CNNs may be powerful enough to learn one step reconstruction, such a network could show signs of overfitting, unless there is vast amounts of training data. In addition, training such networks may require a long time as well as careful fine-tuning steps. It is therefore best to be able to use CNNs for iterative reconstruction approaches.

A simple solution is to train a second CNN which learns to reconstruct from the output of the first CNN. In fact, we can concatenate a new CNN on the output of the previous CNN to build extremely deep networks which iterate between intermediate de-aliasing and the data

consistency reconstruction. We term this a *cascading network*. In fact, one can essentially view this as unfolding the optimisation process of DLMRI. If each CNN expresses the dictionary learning reconstruction step, then the cascading CNN can be seen as a direct extension of DLMRI, where the whole reconstruction pipeline can be optimised from training, as seen in Fig. 4.4. In particular, owing to the forward and back-backpropagation rules defined for the DC layer, all subnetworks can be trained jointly in an end-to-end manner, yielding one large network.

4.5 Data sharing layer

For the case of reconstructing dynamic sequences, the temporal correlation between frames can be exploited as an additional regulariser to further de-alias the undersampled images. For this, we use 3D convolution to learn spatio-temporal features of the input sequence. In addition, we propose incorporating features that could benefit the CNN reconstruction, inspired by *data sharing* approaches [Rie+88], [Ras+95], [ZBF10]: if the change in image content is relatively small for any adjacent frames, then the neighbouring k-space samples along the temporal-axis often capture similar information. In fact, as long as this assumption is valid, for each frame, we can fill the entries using the samples from the adjacent frames to approximate missing k-space samples. Specifically, for each frame t, all frames from $t - n_{adj}$ to $t + n_{adj}$ are considered, filling the missing k-space samples at frame t. If more than one frame within the range contains a sample at the same location, we take the weighted average of the samples. The idea is demonstrated in Fig. 4.2.

An example of data sharing with $n_{adj} = 2$ applied to the Cartesian undersampling is shown in Fig. 4.3(a). As data sharing aggregates the lines in k-space, the resulting images can be seen as a zero-filled reconstruction from a measurement with lower undersampling factor. In practice, however, cardiac sequences contain highly dynamic content around the heart and hence combining the adjacent frames results in data inconsistency around the dynamic region, as illustrated in Fig. 4.3(b,c,d). However, for CNN reconstruction, we can incorporate these images as an

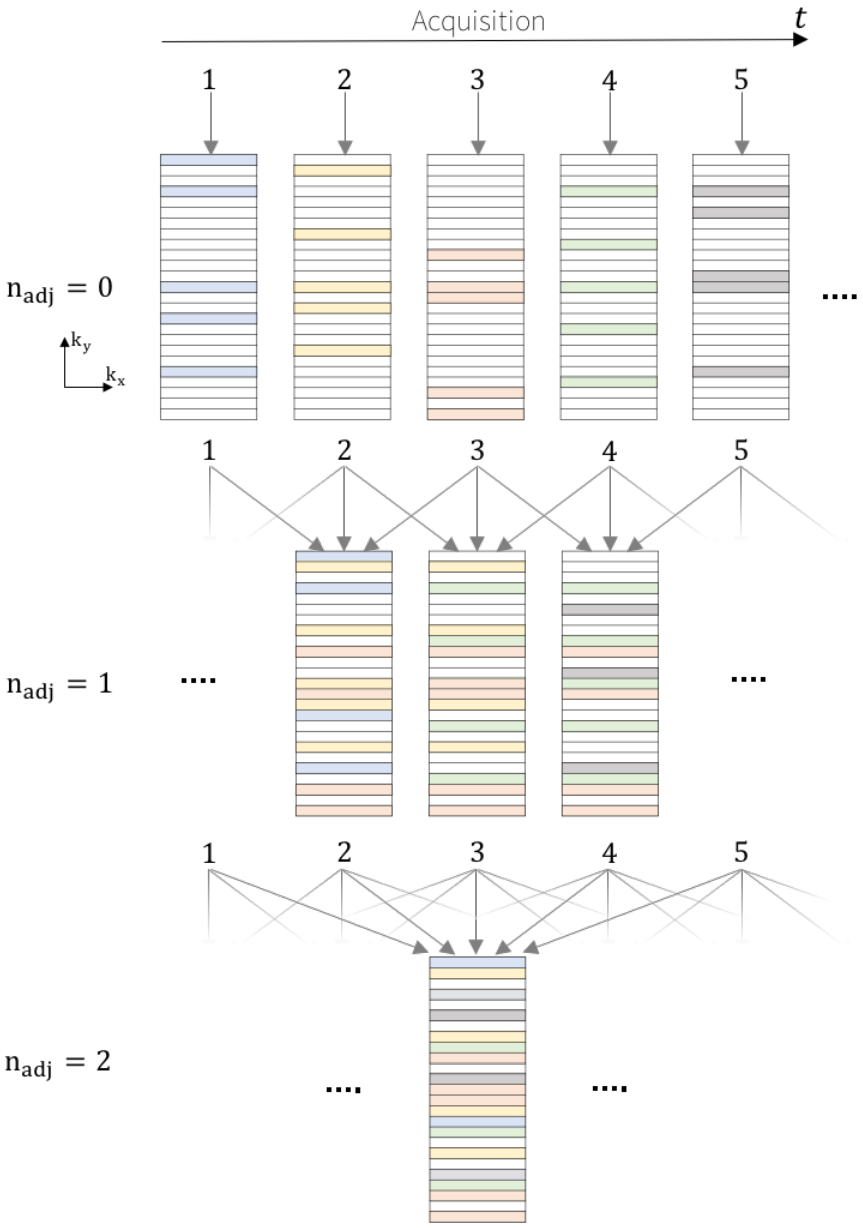

Figure 4.2: The illustration of data sharing approach. The acquired lines, which can be seen as $n_{adj} = 0$, are colour-coded for each time frame. For each n_{adj}, the missing entries in each frame are aggregated using the values from up to $\pm n_{adj}$ neighbouring frames. The overlapped lines are averaged.

extra input to train the network rather than treating them as the final reconstructions. Note that the reduction in the apparent acceleration factor is non-trivial to calculate: if each frame samples 10% of k-space, combining 5 adjacent frames in theory should cover 50%. However, one often relies on variable density sampling, which samples low-frequency terms more often,

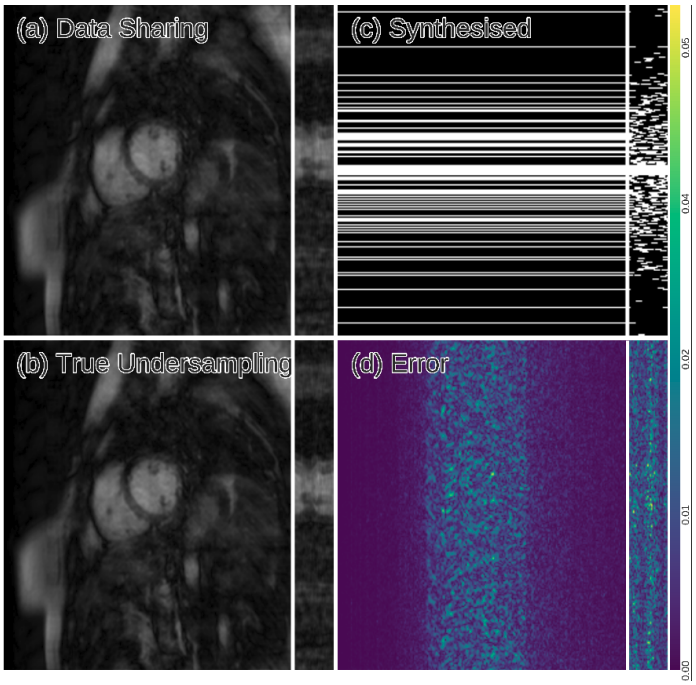

Figure 4.3: The illustration of data sharing approach applied to the image and the mask from Fig.4.1(a,b). In this figure, (a) shows the appearance of the resulting sequence for $n_{adj} = 2$. (b) The entries in k-space that are either acquired or aggregated using the data sharing approach with $n_{adj} = 2$, which conceptually defines a sampling mask. (c) For a comparison, we show the resulting zero-filled reconstruction if (b) were treated as a mask. (d) The error map between the (a) and (b). One can observe their similarity except for the data *inconsistency* of the dynamic content around the heart region. Note that for $n_{adj} = 2$, the obtained image has the appearance similar to acceleration factor around 4 (rather than $12/5 = 2.4$, which is the maximum achievable from 5 frames) due to overlapping lines.

yielding overlapped lines between the adjacent frames. Therefore, the apparent acceleration factor is often much less. As a remedy, regular sampling can be considered. However, regular sampling results in coherent artifact in the image domain, the removal of which is a different problem from the one we address here, which attempts to resolve *incoherent* aliasing patterns. Alternatively, one can perform a sampling trajectory optimisation to reduce the overlapping factor, however, this is out-of-scope for this work and will be investigated in future.

For our network, we implement *data sharing (DS) layers* which take an input image and generate multiple "data-shared" images for a range of n_{adj}. The resulting images are concatenated along the channel-axis and treated as a new input fed into the first convolution layer of the CNNs. Therefore, using the images obtained from data sharing can be interpreted as transforming the problem into joint estimation of aliasing as well as the dynamic motion, where the

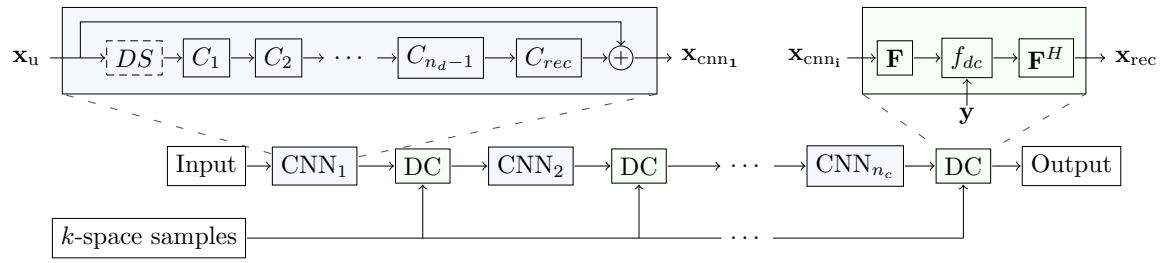

Figure 4.4: A cascade of CNNs. DC denotes the data consistency layer and DS denotes the data sharing layer. The number of convolution layers within each network and the depth of cascade is denoted by n_d and n_c respectively. Note also that DS layer only applies when the input is a sequence of images.

effect of aliasing is considerably smaller. Note that for the cascading network architecture, from the second subnetwork onwards, the input to each subnetwork is no longer "undersampled", but instead contains intermediate predicted values from the previous subnetwork. In this case, we average all the entries from the adjacent frames and update the samples which were not initially acquired. For this work, we allocate equal weight on all adjacent k-space samples, however, in future, more elaborate averaging schemes can be considered.

4.6 Architecture and implementation

Incorporating all the new elements mentioned above, we can devise our cascading network architecture. Our CNN takes in a two-channeled sequence of images $\mathbb{R}^{2N_xN_yN_t}$, where the channels store real and imaginary parts of the zero-filled reconstruction in the image domain. Based on literature, we used the following network architecture for the CNN, illustrated in Fig. 4.4: it has $n_d - 1$ 3D convolution layers C_i, which are all followed by Rectifier Linear Units (ReLU) as a choice of nonlinearity. For each of them, we used a kernel size $k = 3$ [Sze+15] and the number of filters was set to $n_f = 64$. The final layer of the CNN module is a convolution layer C_{rec} with $k = 3$ and $n_f = 2$, which projects the extracted representation back to the image domain. We also used *residual connection* [He+16], which sums the output of the CNN module with its input. Finally, we form a cascading network by using the DC layers interleaved with the CNN reconstruction modules n_c times. For DS layer, we take the input to each subnetwork, generating images for all $n_{adj} \in \{0, 1, \ldots, 5\}$. As aforementioned, the resulting images are

concatenated along the channel-axis and fed to the first convolution layer. We found that this choice of architecture works sufficiently well, however, the parameters were not optimised and there is therefore room for refinement of the results presented. Hence the result is likely to be improved by, for example, incorporating pooling layers and varying the parameters such as kernel size and stride [RFB15], [YK15].

Our model can also be used for 2D image reconstruction by setting $N_t = 1$ and use 2D convolution layers instead, however, data sharing does not apply to 2D reconstruction. For the following experiments, we first explore the network configurations by considering 2D MR image reconstruction. We identify our network by the values of n_c, n_d and the use of data sharing. For example, *D5-C2* means a network with $n_d = 5$, $n_c = 2$ with no data sharing. *D5-C10(S)* corresponds a network with $n_d = 5$, $n_c = 10$ and data sharing.

As mentioned, pixel-wise squared error was used as the objective function. As the proposed architecture is memory-intensive, a small minibatch size is used to train the cascade networks. We used minibatch size 1 for all the experiments but we did not observe any problem with the convergence. We initialised the network weights using He initialisation [He+15]. The Adam optimiser [KB15] was used to train all models, with parameters $\alpha = 10^{-4}, \beta_1 = 0.9$ and $\beta_2 = 0.999$ unless specified. We also added ℓ_2 weight decay of 10^{-7}.

4.7 Experimental results

4.7.1 Setup

Dataset Our method was evaluated using the cardiac MR dataset consisting of 10 fully sampled short-axis cardiac cine MR scans. Each scan contains a single slice SSFP acquisition with 30 temporal frames with a 320×320 mm field of view and 10 mm slice thickness. The data consists of 32-channel data with sampling matrix size 192×190, which was zero-filled to the matrix size 256×256. The raw multi-coil data was reconstructed using SENSE [Pru+99] with no undersampling and retrospective gating. Coil sensitivity maps were normalized to a

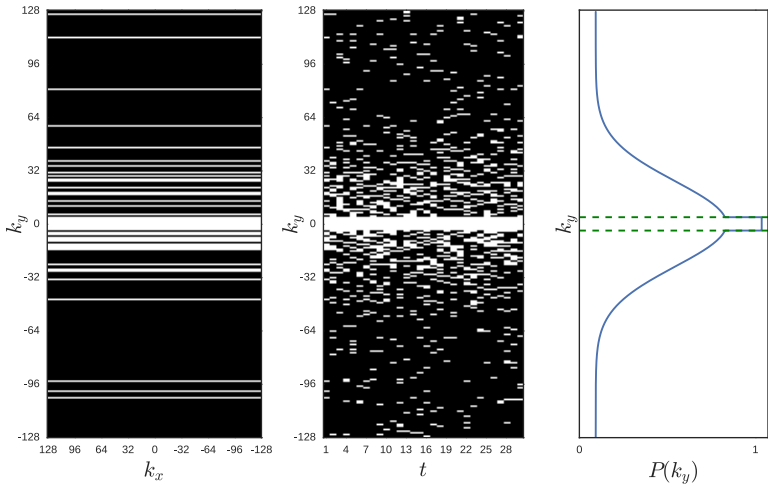

Figure 4.5: The detail of the Cartesian undersampling mask employed in this work. Note that the mask can be seen as a 3D volume indexed by (k_x, k_y, t). For each image frame t, we fully sample along k_x-axis and undersample in k_y direction. We always acquire the 8 central lines and the remaining lines are sampled according to a zero-mean Gaussian distribution with the tail that is marginally offset so it will never reach zero.

body coil image to produce a single complex-valued image set that could be back-transformed to regenerate complex k-space samples or further processed to form final magnitude images. For the following experiments, we perform retrospective undersampling, simulating a practical single-coil acquisition scenario.

Undersampling In this work, we focus on Cartesian undersampling, where one fully samples frequency-encodes (along k_x) and randomly undersamples the phase encodes (along k_y). In addition, we pair consecutive phase encodes, which has been reported to reduce eddy current which is a source of image degredation [BMS05]. For each frame, the eight lowest spatial frequencies are always acquired and other frequencies have a probability of being acquired determined by a zero-mean Gaussian variable density function that is marginally offset, such that the probability of acquisition never reaches zero even at the highest frequencies. An implementation of this approach can be found in [JYK07], and an example of a 2D mask and its effect on the magnitude of a temporal frame is shown in Fig. 4.5. For each experiment, the undersampling rate is fixed and will be stated. For training, the sampling masks were generated on-the-fly to allow the network to learn the differences between potential aliasing artefacts and the underlying signal better. Note that for each acceleration factor *acc*, one can

generate $\binom{k_y}{k_y/acc}$ different masks.

While Cartesian acquisition is the most common protocol in practice and offers straightforward implementation using fast Fourier transform (FFT), other practical sampling strategies such as radial [BUF07] or spiral [Lus+05] could be considered, which achieve greater aliasing incoherence. Nevertheless, they require the use of methods such as nonuniform Fourier transforms and gridding [Fes07] which could propagate interpolation errors.

Data Augmentation Typically, deep learning benefits from large datasets, which are often not available for medical images. Our dataset is relatively small (300 images), however, the literature suggests that it is still possible to train a network by applying appropriate data augmentation strategies [RFB15]. Therefore, we follow that practice and apply data augmentation including rigid transformation and elastic deformation to counter overfitting. Specifically, given each image (or a sequence of images), we randomly apply translation up to ± 20 pixels along x and y-axes, rotation of $[0, 2\pi)$, reflection along x-axis by 50% of chance. Therefore, from rigid transformation alone, we create 0.3 million augmented data per image. Combined with the on-the-fly generation of undersampling masks, we generate very large dataset. For the dynamic scenario, we further added elastic deformation, using the implementation in [S+03], with parameters $\alpha \in [0, 3]$ and $\sigma \in [0.05, 0.1]$, sampled uniformly, as well as reflection along temporal axis. Note that while strong elastic deformation may produce anatomically unrealistic shapes its use is justified as our goal is to train a network which learns to *de-alias* the underlying object in the image, rather than explicitly learning the anatomical shapes.

Evaluation Methodology For the 2D experiments, we split the dataset into training and testing sets including 5 subjects each. Each image frame in the sequence is treated as an individual image, yielding a total of 150 images per set. Note that typically, a portion of training data is treated as a validation set utilised for early-stopping [Bis06], where one halts training if the validation error starts to increase. Initially, we used 3-2-5 split for training, validation and testing. However, even after 3 days of training *cascade* networks, we did not observe any decrease in the validation error. Therefore, we instead included the validation set

in the training to further improve the performance but fix the number of backpropagation to be an order of 10^5, which we empirically found to be sufficient. For the dynamic experiments, we used 7-3 split for training and testing and an order of 10^4 for the number of backpropagation.

To evaluate the performances of the trained networks, we used mean squared error (MSE) as our quantitative measure. The reconstruction signal-to-noise ratio from undersampled data is highly dependent on the imaging data and the undersampling mask. To take this into consideration for fair comparison, we assigned an arbitrary but fixed undersampling mask for each image in the test data, yielding a fixed number of image-mask pairs to be evaluated.

4.7.2 Reconstruction of 2D images

Trade-offs between n_d and n_c

In this experiment we compared two architectures: *D5-C2* ($n_d = 5, n_c = 2$) and *D11-C1* ($n_d = 11, n_c = 1$) to evaluate the benefit of the DC step. The two networks have equivalent depths when the DC layers are viewed as feature extraction layers. However, the former can build deeper features of the image, whereas the latter benefits from the intermediate data consistency step. The undersampling rate was fixed to 3-fold and each network was trained end-to-end for 3×10^5 backpropagations.

The MSE's on the training and test data are shown in Fig. 4.6. Note that a gap between the performance on training and test set may exist by the nature of the dataset (e.g. due to image features, initial level of aliasing, etc.) and therefore it is more informative to study in combination the rate of improvement and the slope at the tail of the curves to assess the overfitting process. Indeed, one can observe that *D11-C1* eventually started to overfit the training data after about 1.2×10^5 backpropagations. As one would expect, since our dataset is small, deep networks can overfit easily. On the other hand, both train and test errors for *D5-C2* were notably lower and had relatively tighter gap, showing better generalisability compared to *D11-C1*. This is suggestively because the architecture employs two data consistency steps and rebuilds the representations at each cascading iteration. This suggests that it is more beneficial

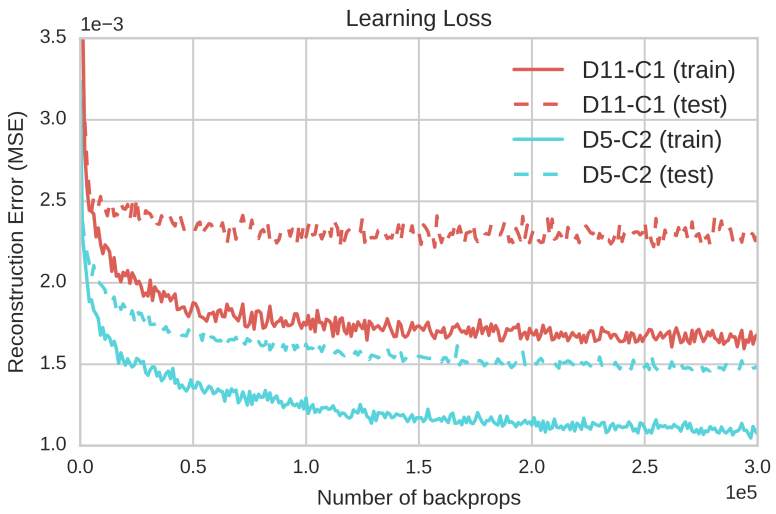

Figure 4.6: A comparison of the networks with and without the intermediate DC step. *D5-C2* shows superior performance over *D11-C1*. In particular, *D5-C2* has considerably lower test error, showing an improved generalization property.

to interleave DC layers projecting the acquired k-space onto intermediate reconstructions with the CNN image reconstruction modules, which appears to help both the reconstruction as well as the generalisation. Nevertheless, there is a considerable gap between train and test data even for *D5-C2*. However, we note from the figure that even after 3×10^5 backpropagations, the test error is still improving. Therefore, although it seems that the network gets more optimised to the features in training data quickly, it still learns features generalisable to test data. Having more training data is likely to accelerate the learning process.

Effect of cascading iterations n_c

In this experiment, we explored how much benefit the network can get by increasing the cascading iteration. We fixed the architectures to have $n_d = 5$, but varied the cascading iteration $n_c \in \{1, 2, 3, 4, 5\}$. For this section, due to time constraints, we trained the networks using a greedy approach: we initialised the cascading net with $n_c = k$ using the net with $n_c = k - 1$ that was already trained. For each n_c, we performed 10^5 backpropagations. Note that the greedy approach leads to a satisfactory solution, however, better results can be achieved with random initialisation, as initialising a network from another networks convergence point can make it more likely that it gets stuck in suboptimal local minima.

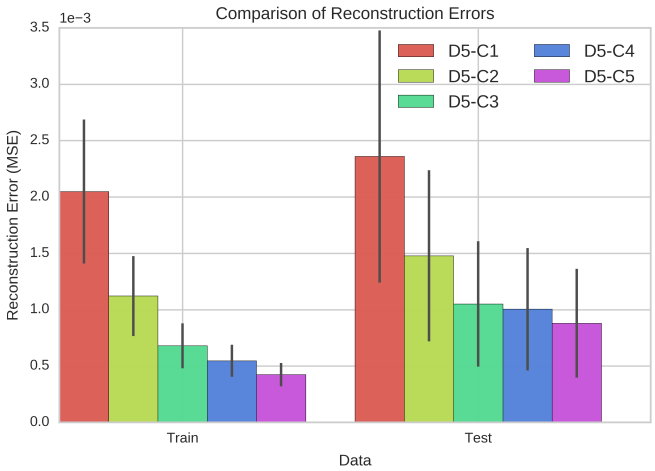

Figure 4.7: The effect of increasing cascading iteration n_c. One can see that the reconstruction error on both training and test data monotonically decreases as n_c increases. However, the rate of improvement is reduced after $n_c = 3$.

Reconstruction errors for each cascading network of different n_c are shown in Fig. 4.7. We observed that while deeper cascading nets tend to overfit more, they still reduced the test error every time. The rate of improvement was reduced after 3 cascading layers, however, we see that the standard deviation of error was also reduced for the deeper models. In the interest of space, we have not shown the resulting images of each *D5-C*n_c but we have observed that increasing n_c resulted in images with more of the subtle image details correctly reconstructed and there was also less noise-like aliasing remaining in the images.

On the other hand, in Fig. 4.8, we show the *intermediate* reconstructions from each subnetwork within *D5-C5* to better understand how the network exploits the iterative nature internally. In general, we see that the cascading net gradually recovers and sharpens the output image. Although the reconstruction error decreased monotonically at each cascading depth, we observed that the output of the fourth subnetwork appears to be more grainy than the output of the preceding subnetwork. This suggests the benefit of the end-to-end training scheme: since we are optimising the whole pipeline of reconstruction, the additional CNNs are internally used to rectify the error caused by the previous CNNs. In this case, the fourth subnetwork appears to counteract over-smoothing in the third subnetwork.

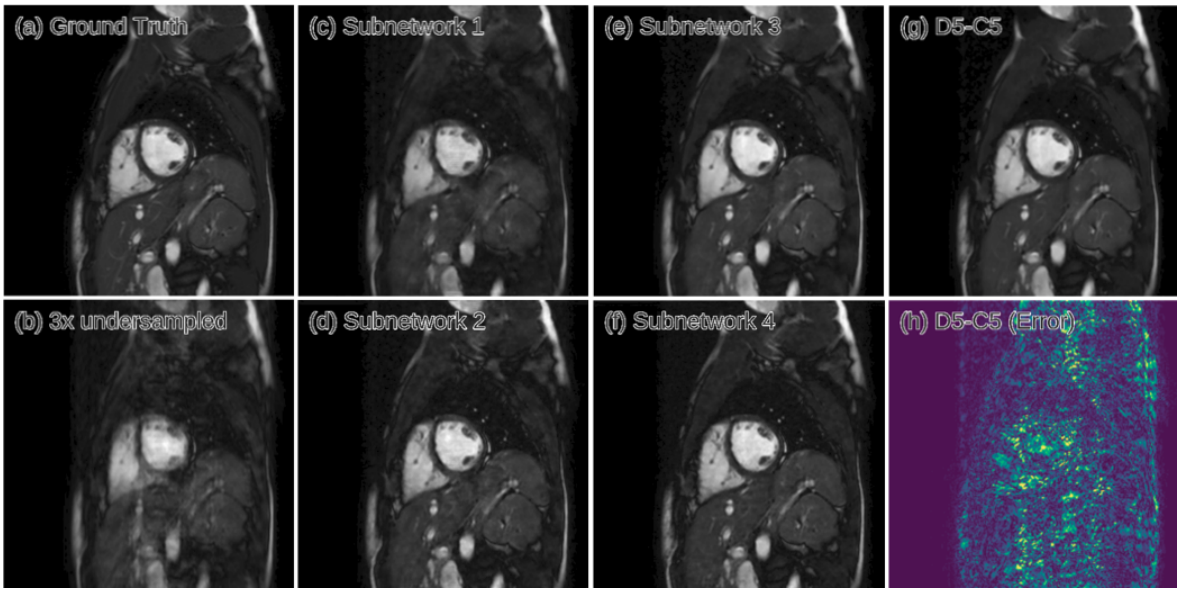

Figure 4.8: 2D reconstruction results of $D5$-$C5$ for one of the test subjects. Here we inspect the intermediate output from each subnetwork in *D5-C5*. (a) Ground truth (b) The input to the network that was 3x undersampled image. The output of (c) first, (d) second, (e) third, (f) fourth cascading subnetwork respectively. (g,h) The final output and the corresponding error. Note that this is not the reconstruction results from the networks in Experiment in 4.7.2.

Comparison with DLMRI

In this experiment, we compared our model with the state-of-the-art dictionary learning-based method, DLMRI, for reconstructing individual 2D cardiac MR images. The comparison was performed for 3-fold and 6-fold acceleration factors.

Models For CNN, we selected the parameters $n_d = 5$, $n_c = 5$. To ensure a fair comparison, we report the aggregated result on the test set from two-way cross-validation (i.e. two iterations of train on five subjects and test on the other five). For each iteration of the cross validation, the network was end-to-end trained using He intialisation [He+15]. For 6-fold undersampling, we initialised the network using the parameters obtained from the trained models from 3-fold acceleration. Each network was trained for 3×10^5 backpropagations, which took one week to train per network on a GeForce TITAN X, however, our manual inspection of the loss curve indicates that the training error plateaued at much early stage, approximately within 3 days.

For DLMRI, we used the implementation from [RB11] with patch size 6×6. Since DLMRI

Table 4.1: The result of 2D reconstruction. DLMRI vs. CNN across 10 scans

	3-fold	6-fold
Models	MSE (SD) $\times 10^{-3}$	MSE (SD) $\times 10^{-3}$
DLMRI	2.12 (1.27)	6.31 (2.95)
CNN (2D)	**0.89** (**0.46**)	**3.42** (**1.65**)

is quite time consuming, in order to obtain the results within a reasonable amount of time, we trained a joint dictionary for all time frames within the subject and reconstructed them in parallel. Note that we did not observe any decrease in performance from this approach. For each subject, we ran 400 iterations and obtained the final reconstruction.

Results The means of the reconstruction errors across 10 subjects are summarised in Table. 4.1. For both 3-fold and 6-fold acceleration, one can see that CNN consistently outperformed DLMRI, and that the standard deviation of the error made by CNN was smaller. The reconstructions from 6-fold acceleration is in Fig. 4.9. Although both methods suffered from significant loss of structures, the CNN was still capable of better preserving the texture than DLMRI (highlighted in red ellipse). On the other hand, DLMRI created block-like artefacts due to over-smoothing. 6x undersampling for these images typically approaches the limit of sparsity-based methods, however, the CNN was able to predict some anatomical details which was not possible by DLMRI. This could be due to the fact that the CNNs has more free parameters to tune with, allowing the network to learn complex but more accurate end-to-end transformations of data.

Comparison of Reconstruction Speed While training the CNN is time consuming, once it is trained, the inference can be done extremely quickly on a GPU. Reconstructing each slice took 23 ± 0.1 milliseconds on a GeForce GTX 1080, which enables real-time applications. To produce the above results, DLMRI took about 6.1 ± 1.3 hours per subject on CPU. Even though we do not have a GPU implementation of DLMRI, it is expected to take longer than 23ms because DLMRI requires dozens of iterations of dictionary learning and sparse coding steps. Using a fixed, pre-trained dictionary could remove this bottleneck in computation although this would

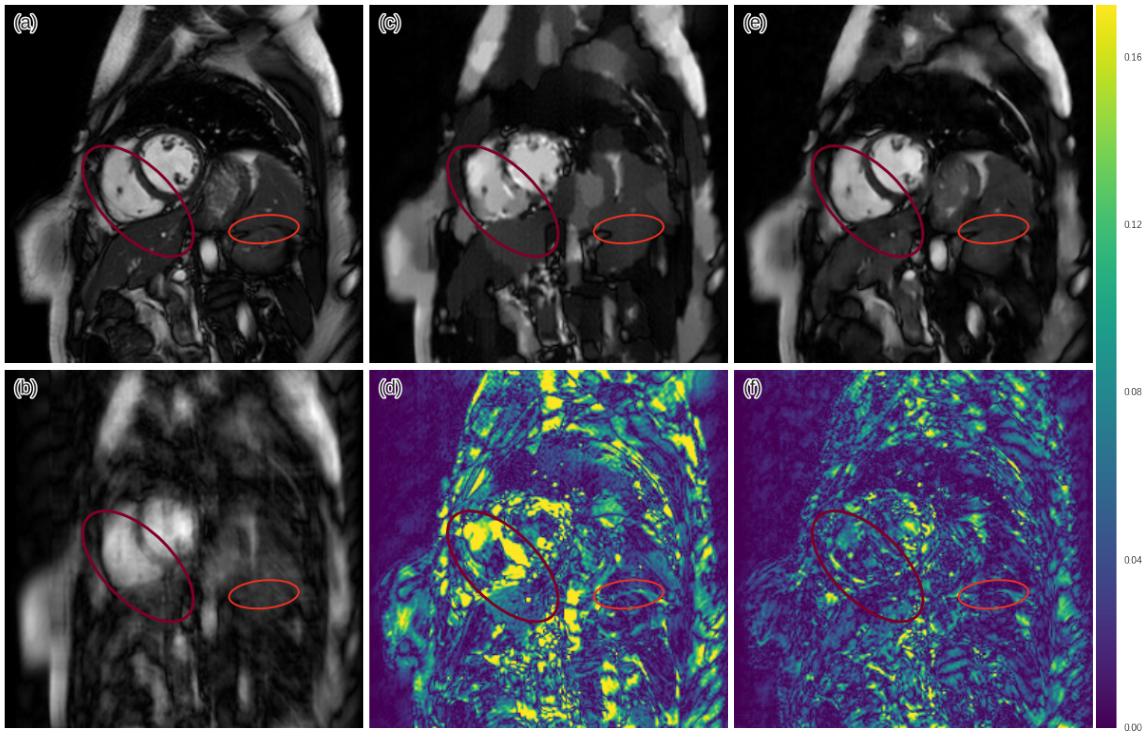

Figure 4.9: The comparison of 2D reconstructions from DLMRI and CNN for test data. (a) The original (b) 6x undersampled (c,d) DLMRI reconstruction and its error map (e,f) CNN reconstruction and its error map. There are larger errors in (d) than (f) and red/orange ellipses highlight the anatomy that was reconstructed by CNN better than DLMRI.

likely be to the detriment of reconstruction quality.

4.7.3 3D experiments

For the following experiments, we split our dataset into training and testing sets containing seven and three subjects respectively. Compared to the 2D case, we have significantly less data. As aforementioned, we applied elastic deformations in addition to rigid transformation to augment the training data input in order to increase the variation of the examples seen by the network. Furthermore, working with a large input is a burden on memory, limiting the size of the network that can be used. To address this, we trained our model on an input size $256 \times N_{patch} \times 30$, where the direction of patch extraction corresponds to the frequency-encoding direction. In this way, we can train the network with the same aliasing patterns while reducing the input size. Note that the extracted patches of an image sequence will have different k-space

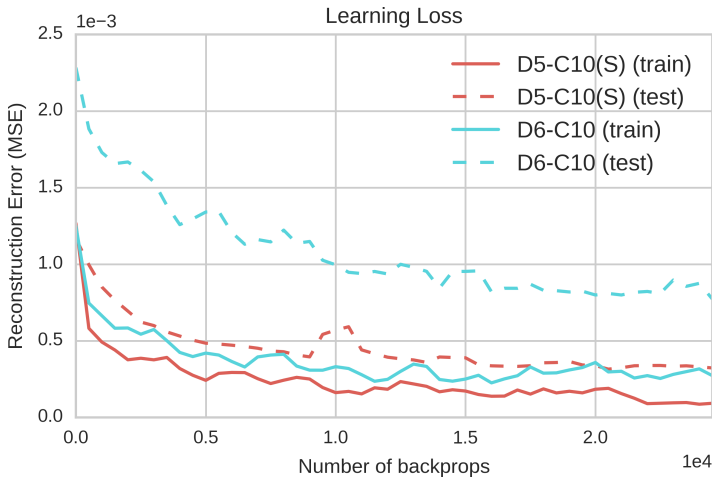

Figure 4.10: The effect of data sharing. The network with data sharing shows superior performance over the other. In particular, it has considerably lower test error, showing an improved generalisation property.

values compared to the original data once the field-of-view (FOV) is reduced. As such, this trick only works for training where the patches can be treated as the new instances of training data. In particular at test time, since only the raw data with full FOV is available, the CNN must also be applied to the entire volume in order to perform data consistency step correctly.

Effect of data sharing

In this experiment, we evaluated the effect of using the features obtained from data sharing. We trained the following two networks: *D5-C10(S)* ($n_d = 5$, $n_c = 10$ with data sharing) and *D6-C10* ($n_d = 6$, $n_c = 10$ without data sharing). In the second network, the data sharing is replaced by an additional convolution layer to account for the additional input. We trained each model to reconstruct the sequences from 9-fold undersampling for 2.5×10^4 backpropagations. Their learning is plotted in Fig. 4.10. We can notice that there is a considerable difference in their errors. The error of the *D5-C10(S)* was smaller for both train and test, suggesting that it was able to learn a strategy to de-alias image that generalises better. Moreover, by using data sharing, the network was able to learn faster. The visualisation of their reconstructions can be found in the following section.

Comparison with state-of-the-art

In this experiment, we compared our model with state-of-the-art methods: DLTG [Cab+14b], kt-SLR [Lin+11] and L+S [OCS15] for reconstructing the dynamic sequence. We compared the results for 3, 6, 9 and 11-fold acceleration factors.

Models For the CNN, we used $n_d = 5$, $n_c = 10$ with data sharing as explained above. We also set the weight decay to 0 as we did not notice any overfitting of the model. Contrary to the 2D case, we trained each network as follows: we first pre-trained the network on various undersampling rates (0-9x) for 5×10^4 backpropagations. Subsequently, each network was fine-tuned for a specific undersampling rate using Adam with learning rate reduced to 5×10^{-5} for 10^4 backpropagations. We performed three way cross validation (where for two iterations we train on 7 subjects then test on 3 subjects, one iteration where we train on 6 subjects and test on 4 subjects) and we aggregated the test errors. The pre-training and the fine tuning stages took approximately 3.5 days and 14 hours respectively using a GeForce GTX 1080. Since the training is time consuming, we did not train the networks longer but we speculate that the network will benefit from further training using lower learning rates. For DLTG, we used the default parameters described in [Cab+14b]. For kt-SLR, we performed grid search to identify the optimal parameters for the data, which were $\mu_1 = 10^{-5}$, $\mu_2 = 10^{-8}$, $\rho = 0.1$. Similarly for L+S, the optimal parameters were $\lambda_L = 0.01$ $\lambda_S = 0.01$.

Result The final reconstruction error is summarised in Fig. 4.11. we see that CNN consistently outperforms state-of-the-art methods for all undersampling factors. For a low acceleration factor (3x undersampling), all methods performed approximately the same, however, for more aggressive undersampling factors, CNN was able to reduce the error by a considerable margin. For aggressive undersampling rates, the performance of kt-SLR and L+S degraded much faster. These methods employ low-rank and simple sparsity constraints. We speculate that they underperformed in this regime because the data is not exactly low-rank (as our temporal dimension is already small) as well as the sparsifying transforms (temporal FFT for L+S

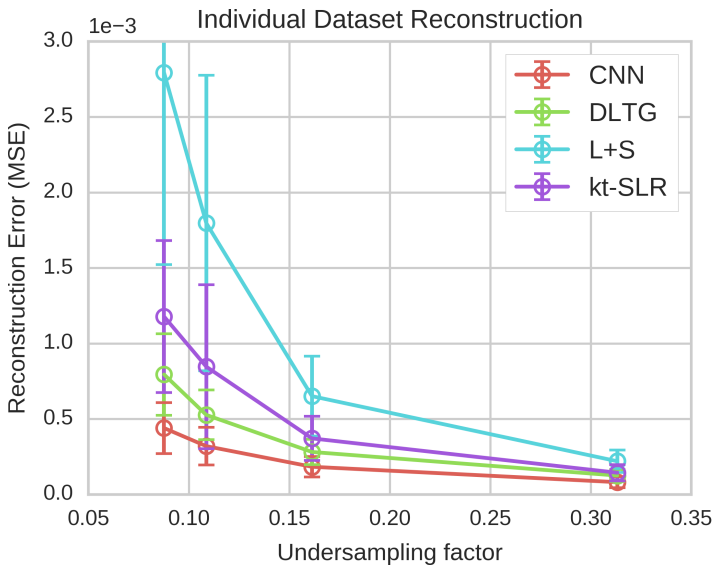

Figure 4.11: The reconstruction errors of CNN vs state-of-the-art methods across 10 subjects for different undersampling rates. Note that we average over the test error from all iterations of cross-validation.

and temporal gradient for kt-SLR) lack adaptability to data compared to CNN and DLTG. The visualisation of reconstruction from 9-fold undersampling is shown in Fig. 4.12, including the reconstruction from the CNN without data sharing and DLTG. The reconstructions of kt-SLR and L+S were omitted as their quantitative error were already much worse. One can see that, as with the 2D case, at aggressive undersampling rate dictionary-learning based method produced blocky artefacts, whereas the CNN methods were capable of reconstructing finer details (indicated in red ellipse). On the other hand, for the CNN without data sharing, one can notice grainy noise-like artefacts. Even though it was able to reconstruct the underlying anatomy more faithfully than DLTG, the overall error was worse. However, this artefact was not present in the images reconstructed by the CNN with data sharing. Although the quantitative result is not shown, CNN without data sharing in fact outperformed DLTG for low acceleration factor (3x) but not for more aggressive undersampling factor. This suggests that when the aliasing is severe, more drastic transformation is required, in which case for CNN to do better, we either need to increase depth, which would increase its computation cost, or increase the training samples. This confirms the importance of data sharing and the necessity to exploit the domain knowledge to simplify the learning problem for the case when the data is limited. Temporal profiles from the reconstructions are shown in Fig. 4.13. Even though

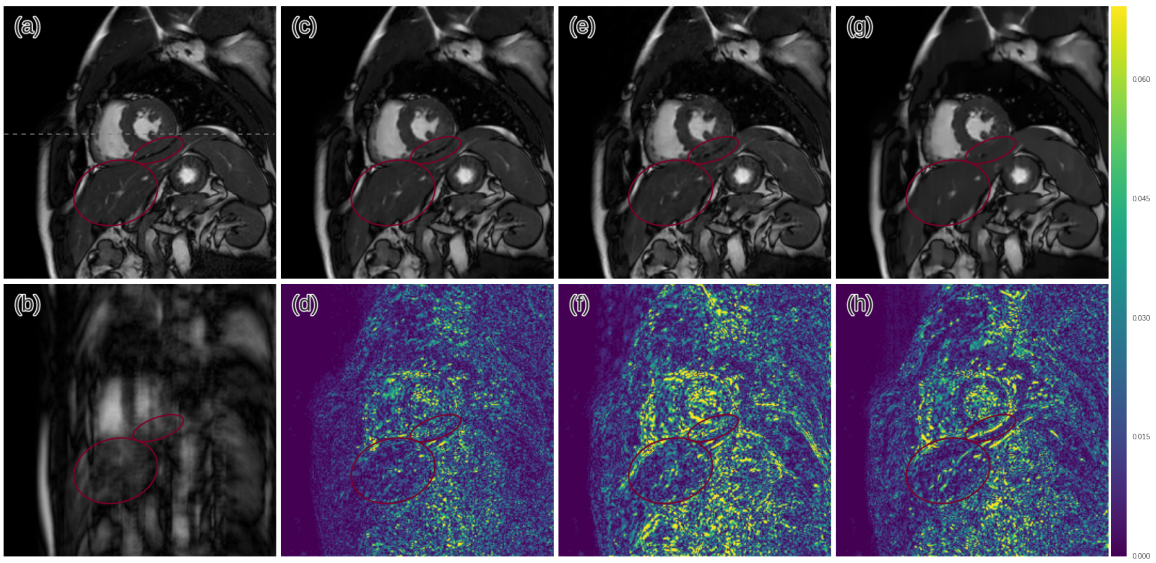

Figure 4.12: The comparison of cardiac MR image sequence reconstructions from DLTG and CNN. Here we show nth slice from one of the test subjects (a) The original (b) 9x undersampled (c,d) CNN with data sharing and its error map (e,f) CNN without data sharing and its error map (g,h) DLTG reconstruction and its error map. Red ellipses highlight the anatomy that was reconstructed by CNN better than DLTG.

the data sharing itself results in data inconsistency in highly dynamic regions, the CNN was able to rectify this internally and reconstructed the correct motion with errors smaller than the other methods. This suggests the CNN's capability solve the joint de-aliasing and implicit estimation of dynamic motion.

Reconstruction with Noise This section analyses the impact of acquisition noise in reconstruction performance. In this experiment we fixed the acceleration factor to be 3 and varied the level of noise in the data. Specifically, we tested for noise power $\sigma^2 \in [10^{-9}, 4 \times 10^{-8}]$. For fully-sampled reconstruction, the noise power is equivalent to peak signal-to-noise (PSNR) values of 41.84 dB and 25.81 dB for 10^{-9} and 4×10^{-8} respectively, where PSNR was calculated as $10 \log_{10}(1/\text{MSE})$. The result is summarised at Fig 4.14, where we aggregate the reconstruction error from all 10 subjects. The input level of noise is indicated by PSNR_f and for consistency, the reconstruction results are also indicated by PSNR (higher the better). For DLTG, we used the value $\lambda = 5 \times 10^{-6}$ as recommended in [Cab+14b]. DLTG showed decent robustness to noise, owing to the nature of underlying K-SVD, which has the effect of sparse

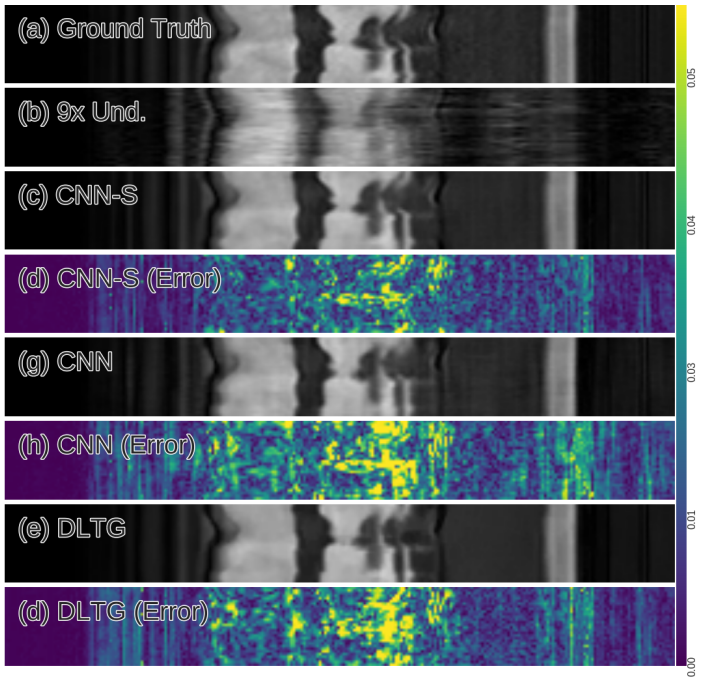

Figure 4.13: The comparison of reconstructions along temporal dimension. Here we extract a 110th slice along y-axis from the previous figure. (a) The original (b) 9x undersampled (c,d) CNN with data sharing and its error map (e,f) CNN without data sharing and its error map (g,h) DLTG reconstruction and its error map.

coding denoising. For kt-SLR and L+S, we used the same parameters as before. They showed some robustness for small noise but they did not perform well in the presence of aggressive noise, as the implementations (and the data consistency step in particular) do not explicitly account for them. Changing such implementation is likely to improve the result.

For CNN, we used the model *D5-C10(S)* as before and tested the following two variations. Firstly, we tested the performance of CNN from the previous section, which were trained in the absence of noise, denoted as *CNN-NAD* (blue curve). It can be seen that for the low level of noise (PSNR $>$ 35 dB), CNN-NAD were able to maintain similar performance as the rest of the methods. However, the performance degraded almost at the same rate as kt-SLR and L+S for the high level of noise. We then trained CNN-NAD to adapt for noise as following. Firstly, we added noise in training data, where we randomly sample the noise power in the range $[10^{-9}, 4 \times 10^{-8}]$. Secondly, we modified our data consistency layers to account for noise. In particular, we initialised λ for each DC layer as $\lambda = q/\sigma = 0.025$ (as in DLTG), made the parameters trainable. We trained the network for 3×10^4 backpropagations and the result is

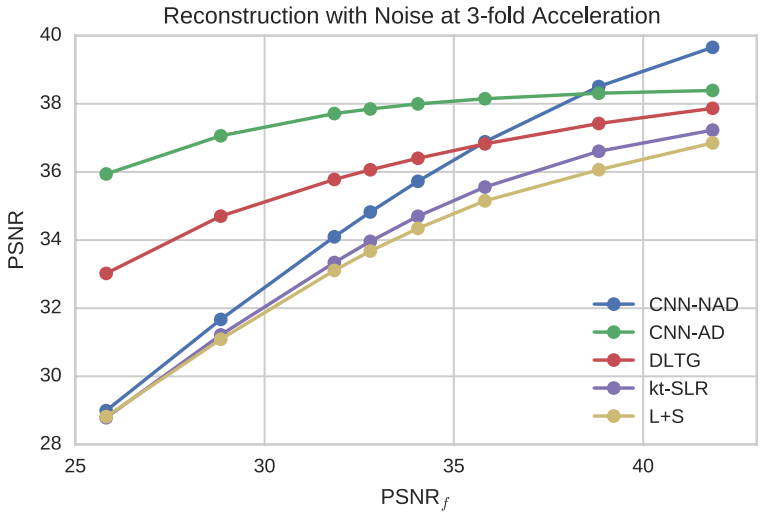

Figure 4.14: The aggregated test error across 10 subjects with injected noise. For different value of input noise power, PSNR$_f$ is shown. The corresponding reconstruction PSNR for CNN-NAD, CNN-AD, DLTG, kt-SLR and L+S are shown.

denoted as *CNN-AD* (green curve). Interestingly, the performance for very small noise ($>$ 38 dB) became worse compared to the original CNN. However, for further acceleration, it showed significant improvement for all level of noise, showing better robustness compared to other methods. We also observed that after fine-tuning, λ was increased to 0.5. This signifies that DLTG and CNN, even though the reconstruction framework shows similarity in terms of the iterative nature, are fundamentally different approaches and the required parameters also vary. Note that since we trained the network for a wide range of noise, the performance is likely to be improved if a narrower range of noise is selected for training. In practice, measuring the level of noise a-priori is non-trivial. However, our CNN showed the adaptability to the pre-specified range which indeed can be simulated in practice.

Reconstruction speed Similar to the 2D case, the DLTG takes 6.6 hours per subject on CPU. For the CNN, each sequence was reconstructed on average $8.21s \pm 0.02s$ on GPU GeForce GTX 1080. This is significantly slower than reconstructing 2D images as introducing a temporal axis greatly increases the computational effort of the convolution operations. Nevertheless, the reconstruction speed of our method is much faster than DLTG and is reasonably fast for offline reconstruction.

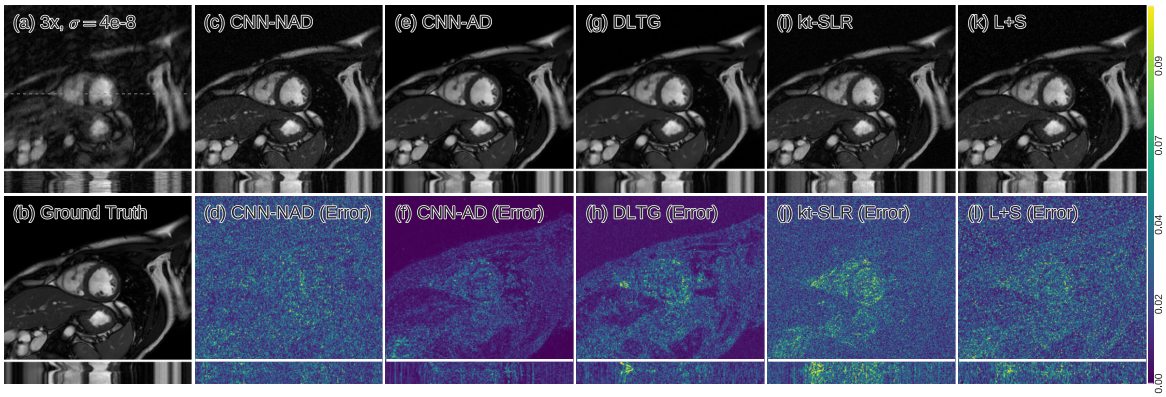

Figure 4.15: The reconstruction with noise $\sigma^2 = 4 \times 10^{-8}$. The aggregated test error across 10 subjects with injected noise. For different value of input noise power, PSNR_f is shown. The corresponding reconstruction PSNR for CNN, finetuned CNN, DLTG, kt-SLR and L+S are shown.

4.7.4 Memory requirement

The memory requirement of the CNNs is based on the number of the network parameters, the number and the sizes of the intermediate activation maps and the space needed for computing the layer operations. The total number of the network parameters is simply given by the sum of all the layer parameters. Each convolution layer has $(k_x k_y k_t n'_f + 1)n_f$ parameters, where k_x, k_y, k_t are the kernel sizes along x, y and t, n'_f and n_f are the number of features of the incoming and current convolution layers respectively and one for the bias. For each DC layer, we also store one parameter for λ. For 2D reconstruction ($k_t = 1$) and *D5-C5* has about 0.6 million parameters, which occupies 2.3MB of the storage assuming single-precision floating point is used ($N_{\text{precision}} = 4$ bytes). For dynamic reconstruction, *D5-C10(S)* has 3.4 million parameters, which occupies about 13.6MB.

At the training stage, more than three times the number of parameters are required for computing the gradient. In addition, all the intermediate activation maps need to be stored to perform the backpropagation efficiently. For the proposed architecture, most of the activation maps are of the convolution layer C_i's; hence, the sum can be roughly estimated by $N_{\text{batch}} N_x N_y N_t N_f n_c (n_d - 1) N_{\text{precision}}$. With the input size $N_{\text{batch}} \times N_x \times N_y \times N_t = 1 \times 256 \times 256 \times 1$, the memory required for the activation maps of *D5-C5* is 335MB. For the dynamic models, the memory requirement further increases by the size of the temporal dimension $N_t = 30$. There-

fore, the aforementioned trick of cropping the images along N_y is necessary to fit the model. For *D5-C10(S)*, with the input size $1 \times 256 \times (256/8) \times 30$, 2.4GB is required for storing the activation maps alone. Finally, to obtain the total memory consumption for the training stage, this value needs to be further multiplied by factors based on the implementation of backpropagation, operations including convolution and FFT as well as any compilation optimisation performed by the library. For example, most implementations of backpropagation require twice the value above accounting for forward- and backward- passes. We report that for our Theano implementation of *D5-C10(S)*, the largest mini-batch size we could fit for the given input size on GeForce GTX 1080 (8GB) was 1.

At the testing stage, the memory requirement is much less because the intermediate activation maps do not need to be stored if only the forward pass needs to be performed. In this case, the memory overhead is only the single largest activation map, which is C_i, scaled by implementation-specific factors. Note that as aforementioned, the patch extraction cannot be used at test time. Nevertheless, we did not observe any problem using *D5-C10(S)* for input size $1 \times 256 \times 256 \times 30$ on GeForce GTX 1080.

4.7.5 Analysis of data consistency layer

In this section, we explore the benefit of using the DC layer in mathematical terms. In particular, we are interested to see how incorporating the DC layer influences the learning process of the network. Firstly, we show that the benefit of the DC layer is that it modifies the objective function so it allocates more weight to errors generated by the entries in k-space that were not initially sampled in $\hat{\mathbf{x}}_u$. Secondly, we show that while a data consistency layer saturates the error through the backward pass, the extent is small so that normalisation layers, such as batch normalisation, are not required for the training.

Let $\mathbf{x}_t$ be the target and $\mathbf{x}_o$ be the output of the DC layer f_L, where

$$\mathbf{x}_o = f_L(\mathbf{x}_{L-1}, \hat{\mathbf{x}}_u) = \mathbf{F}^{-1}(\mathbf{\Lambda}\mathbf{F}\mathbf{x}_{L-1} + \frac{\lambda}{1+\lambda}\hat{\mathbf{x}}_u)$$

and $\mathbf{x}_{l-1}$ is the input to the layer (which is the output of the CNN module). Our objective function is based on an ℓ_2 norm: $\|\mathbf{x}_t - \mathbf{x}_o\|_2^2$, which can be rewritten as follows. First define a diagonal, projection matrix $\mathbf{\Pi}_\Omega$, which has 1 on the ith entry if $i \in \Omega$ and 0 otherwise as well as its additive inverse $\mathbf{\Pi}_{\Omega^c} = \mathbf{I} - \mathbf{\Pi}_\Omega$. These projection matrices are used for separating the contributions of different entries in k-space:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{F}^{-1}(\mathbf{\Pi}_\Omega + \mathbf{\Pi}_{\Omega^c})\hat{\mathbf{x}}_t \\ &= \mathbf{F}^{-1}(\hat{\mathbf{x}}_\Omega + \hat{\mathbf{x}}_{\Omega^c}) \end{aligned} \tag{4.12}$$

We also note that $\mathbf{\Pi}_A \hat{\mathbf{x}}_A = \hat{\mathbf{x}}_A$. Now, the entries in Ω are given by $\hat{\mathbf{x}}_u$, so Eq. (4.12) can be written as:

$$\mathbf{x}_t = \mathbf{F}^{-1}(\hat{\mathbf{x}}_{\Omega^c} + \frac{\boldsymbol{\eta} + \lambda \hat{\mathbf{x}}_u}{1+\lambda})$$

where auxiliary variable $\boldsymbol{\eta}$ was introduced for convenience. In noiseless case, i.e. $\lambda \to \infty$, we have $(\boldsymbol{\eta} + \lambda \hat{\mathbf{x}}_u)/(1+\lambda) \to \hat{\mathbf{x}}_u$. Now the difference in the ℓ_2 norm above can be decomposed as follows:

$$\begin{aligned} \mathbf{x}_t - \mathbf{x}_o &= \mathbf{F}^{-1}(\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_o) \\ &= \mathbf{F}^{-1}\left((\hat{\mathbf{x}}_{\Omega^c} + \frac{\boldsymbol{\eta} + \lambda \hat{\mathbf{x}}_u}{1+\lambda}) - (\Lambda \mathbf{F}\mathbf{x}_{L-1} + \frac{\lambda}{1+\lambda}\hat{\mathbf{x}}_u)\right) \\ &= \mathbf{F}^{-1}\left(\mathbf{\Pi}_{\Omega^c}[\hat{\mathbf{x}}_{\Omega^c}] + \mathbf{\Pi}_\Omega[\frac{\boldsymbol{\eta}}{1+\lambda}] - (\mathbf{\Pi}_{\Omega^c} + \frac{1}{1+\lambda}\mathbf{\Pi}_\Omega)[\mathbf{F}\mathbf{x}_{L-1}]\right) \\ &= \mathbf{F}^{-1}\underbrace{\mathbf{\Pi}_{\Omega^c}[\hat{\mathbf{x}}_{\Omega^c} - \mathbf{F}\mathbf{x}_{L-1}]}_{(*)} - \mathbf{F}^{-1}\underbrace{\frac{1}{1+\lambda}\mathbf{\Pi}_\Omega[\boldsymbol{\eta} - \mathbf{F}\mathbf{x}_{L-1}]}_{(**)} \end{aligned} \tag{4.13}$$

We note that (*) corresponds to the difference of variables where no k-space values have been sampled, whereas (**) aggregates the error for the entries in Ω. In the noiseless setting, i.e. $\lambda \to \infty$, we have (**) $\to 0$, indicating that in such case, we only consider the error from the entries in Ω^c. This intuitively makes sense as initial measurements do not contribute to the final error. In general, we can see that the effect of the data consistency layer is that it allocates

more importance on the error that comes from k-space entries that were not sampled initially.

Now we turn our consideration to its effect on the backpropagation. Recall that the Jacobian of the layer is given by: $\mathbf{F}^{-1}\mathbf{\Lambda}\mathbf{F}$. the DFT matrices are orthogonal and hence preserves the matrix norm, however, $\mathbf{\Lambda}$ clearly attenuates the entries in Ω. Therefore we have:

$$\begin{aligned}\|\mathbf{F}^{-1}\mathbf{\Lambda}\mathbf{F}\mathbf{x}\|_2^2 &= (\mathbf{F}^{-1}\mathbf{\Lambda}\mathbf{F}\mathbf{x})^H(\mathbf{F}^{-1}\mathbf{\Lambda}\mathbf{F}\mathbf{x}) = \hat{\mathbf{x}}^H\mathbf{\Lambda}^H\mathbf{\Lambda}\hat{\mathbf{x}} \\ &= \|\mathbf{\Lambda}\hat{\mathbf{x}}\|_2^2 \\ &\leq \|\hat{\mathbf{x}}\|_2^2 = \|\mathbf{F}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2\end{aligned} \tag{4.14}$$

Notice however, we can write the Jacobian of the error $\mathbf{e} = \mathbf{x}_t - \mathbf{x}_o$ in k-space as follows:

$$\boldsymbol{e} = \mathbf{e}_{\mathbf{\Omega}} + \mathbf{e}_{\mathbf{\Omega^c}} = \mathbf{F}^{-1}\hat{\boldsymbol{e}}_\Omega + \mathbf{F}^{-1}\hat{\boldsymbol{e}}_{\Omega^c}$$

where the scaling constant was omitted for simplicity. Backpropagating through the data consistency layer gives us:

$$\begin{aligned}\mathbf{F}^{-1}\mathbf{\Lambda}\mathbf{F}\boldsymbol{e} &= \mathbf{F}^{-1}(\mathbf{\Pi}_{\Omega^c} + \frac{1}{1+\lambda}\mathbf{\Pi}_\Omega)\mathbf{F}\boldsymbol{e} \\ &= \mathbf{F}^{-1}(\mathbf{\Pi}_{\Omega^c} + \frac{1}{1+\lambda}\mathbf{\Pi}_\Omega)(\hat{\boldsymbol{e}}_{\Omega^c} + \hat{\boldsymbol{e}}_\Omega) \\ &= \mathbf{F}^{-1}\hat{\boldsymbol{e}}_{\Omega^c} + \frac{1}{1+\lambda}\mathbf{F}^{-1}\hat{\boldsymbol{e}}_\Omega \\ &= \boldsymbol{e}_{\Omega^c} + \frac{1}{1+\lambda}\boldsymbol{e}_\Omega\end{aligned} \tag{4.15}$$

The layer inevitably reduces the magnitude of the error, however, for the noiseless case, we have $\boldsymbol{e}_\Omega/(1+\lambda) \to 0$ so the Jacobian reduces to an identity map. With noise, the error from Ω is attenuated. Nevertheless, this makes sense as $\boldsymbol{e}_\Omega$ is contributed by $\hat{\mathbf{x}}_u$, i.e. changing the network parameter $\boldsymbol{\theta}$ in the direction of minimising $\mathbf{e}_\Omega$ gives relatively small improvement compared to updating the entries in Ω^c. In the pathological case where $\lambda \to 0$ then the Jacobian of the data consistency layer again behaves as an identity map, which essentially says there is no data

consistency being imposed.

4.8 Discussion and conclusion

In this work, we evaluated the applicability of CNNs for the challenge of reconstructing undersampled cardiac MR image data. The experiments presented show that using a network with interleaved data consistency stages, it is feasible to obtain a model which can reconstruct images well. The CS and low-rank framework offers a mathematical guarantee for the signal recovery, which makes the approach appealing in theory as well as in practice even though the required sparsity cannot generally be genuinely achieved in medical imaging. However, even though this is not the case for CNNs, we have empirically shown that a CNN-based approach can outperform them. In addition, at very aggressive undersampling rates, the CNN method was capable of reconstructing most of the anatomical structures more accurately based on the learnt priors, while classical methods do not guarantee such behaviour.

Note that remarkably, we were able to train the CNN on the small dataset. We used several strategies to alleviate the issue of overfitting: firstly, as we employed the iterative architecture, each subnetwork has relatively small receptive field. As a result, the network can only perform local transformations. Secondly, we applied intensive data augmentation so the network constantly sees a variation of the input, which makes it more difficult to overfit to any specific patterns. However, we speculate that given more training data, we can drop the data augmentation and let the network learn coarse features by incorporating, for example, dilated or strided convolution, which could further improve the performance.

It is important to note that in the experiments presented the data was produced by retrospective undersampling of back transformed complex images (equivalent to single-coil data) obtained through an original SENSE reconstruction. Although the application of CNN reconstruction needs to be investigated in the more practical scenario of full array coil data from parallel MR, the results presented show a great potential to apply deep learning for MR reconstruction. The additional richness of array coil data has the potential to further improve performance,

although it will also add considerable complexity to the required CNN architecture.

In this work, we were able to show that the network can be trained using arbitrary Cartesian undersampling masks of fixed sampling rate rather than selecting a fixed number of undersampling masks for training and testing. In addition, we were able to pre-train the network on various undersampling rates before fine-tuning the network. This suggests that the network was capable of learning a generic strategy to de-alias the images. A further investigation should consider how tolerant the network is for different undersampling patterns such as radial and spiral trajectories. As these trajectories provide different properties of aliasing artefacts, a further validation is appropriate to determine the flexibility of our approach. However, radial sampling naturally fits well with the data sharing framework and therefore can be expected to push the performance of the network further. The data sharing approach may also make it feasible to adopt regular undersampling patterns which are intrinsically more efficient. Another interesting direction would be to jointly optimise the undersampling mask using the learning framework.

Although CNNs can only learn local representations which should not affect global structure, it remains to be determined how the CNN approach operates when there is a pathology present in images, or other more variable content. We have performed a cross-validation study to ensure that the network can handle unseen data acquired through the same acquisition protocol. Generalisation properties must be evaluated carefully on a larger dataset. However, CNNs are flexible in a way such that one can incorporate application specific priors to their objective functions to allocate more importance to preserving features of interest in the reconstruction, provided that such expert knowledge is available at training time. For example, analysis of cardiac images in clinical settings often employs segmentation and/or registration. Multi-task learning is a promising approach to further improve the utility of CNN-based MR reconstructions.

To conclude, mainly two things can be noted from this chapter. Firstly, the network architecture was by no-mean optimal. In a way, it was a brute-force unrolling of the traditional method. In the next chapter, we investigate how this data representation can be made more optimal by

changing the architecture. Secondly, as suggested above, deep learning based reconstruction is beneficial over traditional approach because it is also flexible framework to harness data. In particular, we can try to extend our modelling to do direct estimation of parameters, which is investigated in Chapter 6.

Chapter 5

Recurrent neural network for dynamic MRI reconstruction

This section is based on the following publication:[1]

- Qin, C.[†], **Schlemper, J.**[†], Caballero, J., Price, A. N., Hajnal, J. V., Rueckert, D. (2018). Convolutional recurrent neural networks for dynamic MR image reconstruction. IEEE transactions on medical imaging, 38(1), 280-290.

5.1 Introduction

Dynamic MRI is a non-invasive imaging technique which attempts to reveal both spatial and temporal profiles of the underlying anatomy, which has a variety of applications such as cardiovascular imaging and perfusion imaging. As aforementioned in Chapter 2, the acquisition speed is fundamentally limited due to both hardware and physiological constraints as well as the requirement to satisfy the Nyquist sampling rate. Long acquisition times are not only a burden for patients but also make MRI susceptible to motion artefacts.

[1] [†] the authors contributed equally.

In Chapter 4, we first surveyed the state-of-the-art compressed sensing approaches which were proposed in hopes to accelerate the image acquisition time. We then pointed out that the compressed sensing is often associated with its own difficulty to optimise the regularisation terms as well as is limited as it does not harness available data. We argued that deep learning is a natural candidate that can implicitly exploit the learnt representation based on large-scale data. In particular, we proposed a deep cascade of CNN's for accelerated MR image reconstruction. We showed that the proposed method outperformed the state-of-the-art compressed sensing methods including dictionary learning and low-rank approaches. However, the proposed architecture was naive in the sense it was a brute-force unrolling of optimisation similar to that of dictionary learning optimisation. In this chapter, the network architecture is first formalised using *variable-splitting* scheme. We then depart from this framework and present a more general approach for the reconstruction process.

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

To the best of our knowledge, this is the first work applying RNNs for dynamic MRI reconstruction. The contributions of this work are the following: Firstly, we view the optimisation problem of dynamic data as a recurrent network and describe a novel CRNN architecture which simul-

taneously incorporates the recurrence existing in both temporal and iteration sequential steps. Secondly, we demonstrate that the proposed method shows promising results and improves upon the current state-of-the-art dynamic MR reconstruction methods both in reconstruction accuracy and speed. Finally, we compare our architecture to 3D CNN which does not impose the recurrent structure. We show that the proposed method outperforms the CNN at different undersampling rates and speed, while requiring significantly fewer parameters.

5.2 Related work

One of the main challenges associated with recovering an uncorrupted image is that both the undersampling strategy and a-priori knowledge of appropriate properties of the image need to be taken into account. Methods like k-t BLAST and k-t SENSE [TBP03] take advantage of a-priori information about the x-f support obtained from the training data set in order to prune a reconstruction to optimally reduce aliasing. An alternative popular approach is to exploit temporal redundancy to unravel from the aliasing by using CS approaches [JYK07; Cab+14b] or CS combined with low-rank approaches [Lin+11; OCS15]. The class of methods which employ CS to the MRI reconstruction is termed as CS-MRI [Lus+08]. They assume that the image to be reconstructed has a sparse representation in a certain transform domain, and they need to balance sparsity in the transform domain against consistency with the acquired undersampled k-space data. For instance, an example of successful methods enforcing sparsity in x-f domain is k-t FOCUSS [JYK07]. A low rank and sparse reconstruction scheme (k-t SLR) [Lin+11] introduces non-convex spectral norms and uses a spatio-temporal total variation norm in recovering the dynamic signal matrix. Dictionary learning approaches were also proposed to train an over-complete basis of atoms to optimally sparsify spatio-temporal data [Cab+14b]. These methods offer great potential for accelerated imaging, however, they often impose strong assumptions on the underlying data, requiring nontrivial manual adjustments of hyperparameters depending on the application. In addition, it has been observed that these methods tend to result in blocky [Ham+18] and unnatural reconstructions, and their reconstruction speed is often slow. Furthermore, these methods are not able to exploit the prior knowledge that can

be learnt from the vast number of MRI exams routinely performed, which should be helpful to further guide the reconstruction process.

Recently, deep learning-based MR reconstruction has gained popularity due to its promising results for solving inverse and compressed sensing problems. In particular, two paradigms have emerged: the first class of approaches proposes to use convolutional neural networks (CNNs) to learn an end-to-end mapping, where architectures such as SRCNN [Don+14] or U-net [RFB15] are often chosen for MR image reconstruction [LYY17; HYY17; Wan+16; Wan+17b]. The second class of approaches attempts to make each stage of iterative optimisation learnable by unrolling the end-to-end pipeline into a deep network [Ham+18; AO18; S+16; Sch+17; AO17]. For instance, Hammernik et al. [Ham+18] introduced a trainable formulation for accelerated parallel imaging (PI) based MRI reconstruction termed variational network, which embedded a CS concept within a deep learning approach. ADMM-Net [S+16] was proposed by reformulating an alternating direction method of multipliers (ADMM) algorithm to a deep network, where each stage of the architecture corresponds to an iteration in the ADMM algorithm. In Chapter 4, we proposed a cascade network which simulated the iterative reconstruction of dictionary learning-based methods for dynamic MR reconstructions [Sch+17; Sch+18a]. Most approaches so far have focused on 2D images, whereas only a few approaches exist for dynamic MR reconstruction [Sch+18a; BRE17]. While they show promising results, the optimal architecture, training scheme and configuration spaces are yet to be fully explored.

More recently, several deep learning methods on 2D MR image reconstruction were proposed [Ham+18; AMJ17; Mar+17], which share similar idea with our proposed method that integrates data fidelity term and regularisation term into a single deep network so that to enable the end-to-end training. In contrast to these methods which use shared parameters over iterations, as we will show, our architecture integrates hidden connections over optimisation iterations to propagate learnt representations across both iteration and time, whereas such information is discarded in the other methods. Such proposed architecture enables the information used for the reconstruction at each iteration to be shared across all stages of the reconstruction process, aiming for an iterative algorithm that can fully benefit from information extracted at all processing stages. As to the nature of the proposed RNN units, previous work involving RNNs

only updated the hidden state of the recurrent connection with a fixed input [Gre+15; LH15; KWW16], while the proposed architecture progressively updates the input as the optimisation iteration increases. In addition, previous work only modelled the recurrence of iteration *or* time [HWW15] exclusively, whereas the proposed method jointly exploits both dimensions, yielding a unique architecture suitable for the dynamic reconstruction problem.

5.3 Problem formulation

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

$$\underset{\mathbf{x}}{\operatorname{argmin}} \quad \mathcal{R}(\mathbf{x}) + \lambda \|\mathbf{y} - \mathbf{F}_u \mathbf{x}\|_2^2 \tag{5.1}$$

Here $\mathbf{F}_u$ is an undersampling Fourier encoding matrix, $\mathcal{R}$ expresses regularisation terms on $\mathbf{x}$ and λ allows the adjustment of data fidelity based on the noise level of the acquired measurements $\mathbf{y}$. For CS and low-rank based approaches, the regularisation terms $\mathcal{R}$ often employed are ℓ_0 or ℓ_1 norms in the sparsifying domain of $\mathbf{x}$ as well as the rank or nuclear norm of $\mathbf{x}$ respectively. In general, Eq. 5.1 is a non-convex function and hence, the variable splitting technique is usually adopted to decouple the fidelity term and the regularisation term. By introducing an auxiliary variable $\mathbf{z}$ that is constrained to be equal to $\mathbf{x}$, Eq. 5.1 can be reformulated to minimise the following cost function via the penalty method:

$$\underset{\mathbf{x},\mathbf{z}}{\operatorname{argmin}} \; \mathcal{R}(\mathbf{z}) + \lambda \|\mathbf{y} - \mathbf{F}_u \mathbf{x}\|_2^2 + \mu \|\mathbf{x} - \mathbf{z}\|_2^2 \tag{5.2}$$

where μ is a penalty parameter. By applying alternate minimisation over $\mathbf{x}$ and $\mathbf{z}$, Eq. 5.2 can be solved via the following iterative procedures:

$$\mathbf{z}^{(i+1)} = \underset{\mathbf{z}}{\operatorname{argmin}}\ \mathcal{R}(\mathbf{z}) + \mu\|\mathbf{x}^{(i)} - \mathbf{z}\|_2^2 \tag{5.3a}$$

$$\mathbf{x}^{(i+1)} = \underset{\mathbf{x}}{\operatorname{argmin}}\ \lambda\|\mathbf{y} - \mathbf{F}_u\mathbf{x}\|_2^2 + \mu\|\mathbf{x} - \mathbf{z}^{(i+1)}\|_2^2 \tag{5.3b}$$

where $\mathbf{x}^{(0)} = \mathbf{x}_u = \mathbf{F}_u^H y$ is the zero-filled reconstruction taken as an initialisation and $\mathbf{z}$ can be seen as an intermediate state of the optimisation process. For MRI reconstruction, Eq. 5.3b is *data consistency* (DC) step presented in the previous chapter, which admits a closed-form solution [Sch+17]:

$$\begin{aligned}
&\mathbf{x}^{(i+1)} = \mathrm{DC}(\mathbf{z}^{(i)}; \mathbf{y}, \lambda_0, \Omega) = \mathbf{F}^H\mathbf{\Lambda}\mathbf{F}\mathbf{z}^{(i)} + \tfrac{\lambda_0}{1+\lambda_0}\mathbf{F}_u^H\mathbf{y}, \\
&\mathbf{\Lambda}_{kk} = \begin{cases} 1 & \text{if } k \notin \Omega \\ \frac{1}{1+\lambda_0} & \text{if } k \in \Omega \end{cases}
\end{aligned} \tag{5.4}$$

in which $\mathbf{F}$ is the full Fourier encoding matrix (a discrete Fourier transform in this case), $\lambda_0 = \lambda/\mu$ is a ratio of regularisation parameters from Eq. 5.4, Ω is an index set of the acquired k-space samples and Λ is a diagonal matrix. Please refer to [Sch+17] for more details of formulating Eq. 5.4 as a data consistency layer in a neural network. Eq. 5.3a is the proximal operator of the prior $\mathcal{R}$, and instead of explicitly determining the form of the regularisation term, we propose to directly learn the proximal operator by using a convolutional recurrent neural network (CRNN).

Previous deep learning approaches such as Deep-ADMM net [S+16] and deep cascade of CNN's [Sch+17] unroll the traditional optimisation algorithm. Hence, their models learn a sequence of transition $\mathbf{x}^{(0)} \to \mathbf{z}^{(1)} \to \mathbf{x}^{(1)} \to \cdots \to \mathbf{z}^{(N)} \to \mathbf{x}^{(N)}$ to reconstruct the image, where each state transition at stage (i) is an operation such as convolutions independently parameterised by $\boldsymbol{\theta}$, nonlinearities or a data consistency step. However, since the network implicitly learns some form of proximal operator at each iteration, it may be redundant to individually parameterise each step. In our formulation, we model each optimisation stage (i) as a learnt, *recurrent*,

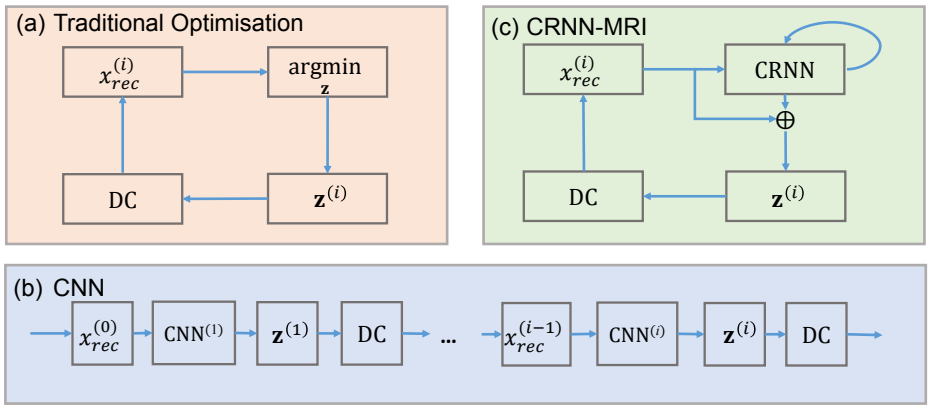

Figure 5.1: (a) Traditional optimisation algorithm using variable splitting and alternate minimisation approach, (b) the optimisation unrolled into a deep convolutional network incorporating the data consistency step, and (c) the proposed architecture which models optimisation recurrence.

forward encoding step $f_i(\mathbf{x}^{(i-1)}, \mathbf{z}^{(i-1)}; \boldsymbol{\theta}, \mathbf{y}, \lambda, \Omega)$. The difference is that now we use one model which performs proximal operator, however, it also allows itself to propagate information across iteration, making it adaptable for the changes across the optimisation steps. The detail will be discussed in the following section. The different strategies are illustrated in Fig 5.1.

5.4 CRNN for MRI reconstruction

RNN is a class of neural networks that makes use of sequential information to process sequences of inputs. They maintain an internal state of the network acting as a "memory", which allows RNNs to naturally lend themselves to the processing of sequential data. Inspired by iterative optimisation schemes of Eq. 5.3, we propose a novel convolutional RNN (CRNN) network. In the most general scope, our neural encoding model is defined as follows,

$$\mathbf{x}_{rec} = f_N(f_{N-1}(\cdots(f_1(\mathbf{x}_u)))), \tag{5.5}$$

in which $\mathbf{x}_{rec}$ denotes the prediction of the network, $\mathbf{x}_u$ is the sequence of undersampled images with length T and also the input of the network, $f_i(\mathbf{x}_u; \boldsymbol{\theta}, \lambda, \Omega)$ is the network function for each iteration of optimisation step, and N is the number of iterations. We can compactly represent

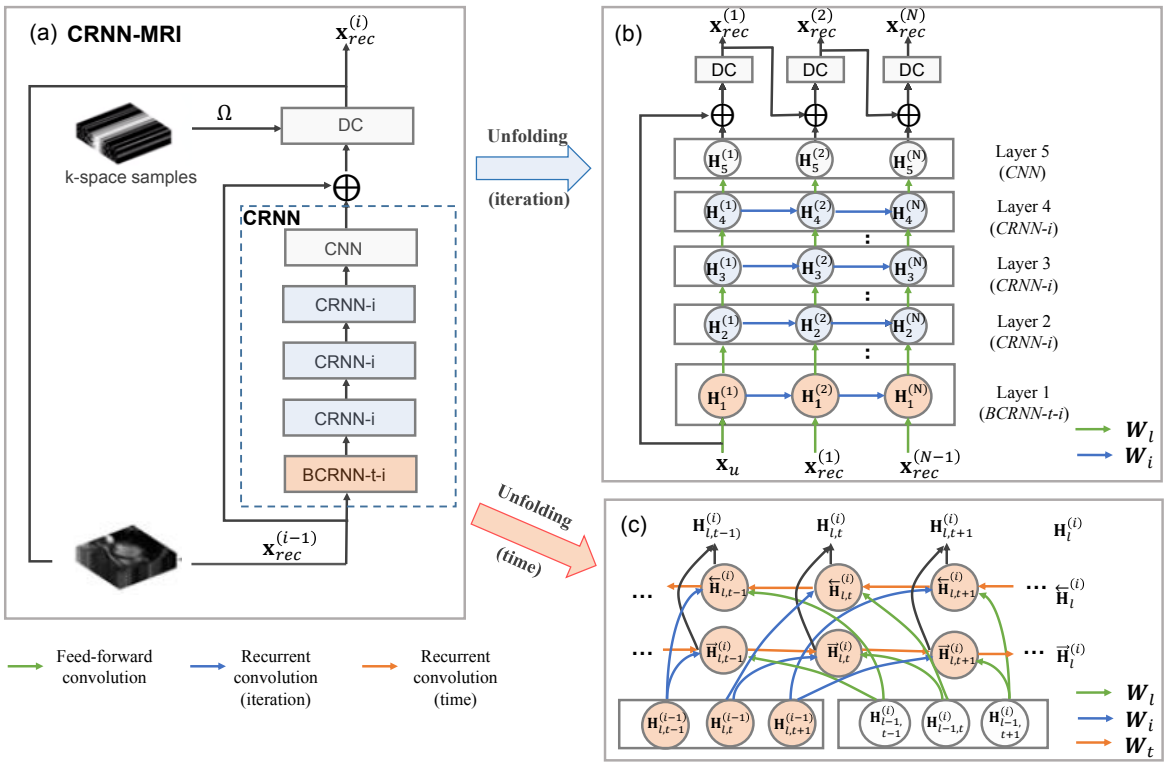

Figure 5.2: (a) The overall architecture of proposed CRNN-MRI network for MRI reconstruction. (b) The structure of the proposed network when unfolded over iterations, in which $\mathbf{x}_{rec}^{(0)} = \mathbf{x}_u$. (c) The structure of BCRNN-t-i layer when unfolded over the time sequence. The green arrows indicate feed-forward convolutions which are denoted by W_l. The blue arrows (W_i) and red arrows (W_t) indicate recurrent convolutions over iterations and the time sequence respectively. For simplicity, we use a single notation to denote weights for these convolutions at different layers. However, in the implementation, the weights are independent across layers.

a single iteration f_i of our network as follows:

$$\mathbf{x}_{rnn}^{(i)} = \mathbf{x}_{rec}^{(i-1)} + \text{CRNN}(\mathbf{x}_{rec}^{(i-1)}), \tag{5.6a}$$

$$\mathbf{x}_{rec}^{(i)} = \text{DC}(\mathbf{x}_{rnn}^{(i)}; \mathbf{y}, \lambda_0, \Omega), \tag{5.6b}$$

where CRNN is a learnable block explained hereafter, DC is the data consistency step treated as a network layer, $\mathbf{x}_{rec}^{(i)}$ is the progressive reconstruction of the undersampled image $\mathbf{x}_u$ at iteration i with $\mathbf{x}_{rec}^{(0)} = \mathbf{x}_u$, $\mathbf{x}_{rnn}^{(i)}$ is the intermediate reconstruction image before the DC layer, and $\mathbf{y}$ is the acquired k-space samples. Note that the variables $\mathbf{x}_{rec}, \mathbf{x}_{rnn}$ are analogous to $\mathbf{x}, \mathbf{z}$ in Eq. 5.3 respectively. Here, we use CRNN to encode the update step, which can be seen as one step of a gradient descent in the sense of objective minimisation, or a more general

approximation function regressing the difference $\mathbf{z}^{(i+1)} - \mathbf{x}^{(i)}$, i.e. the distance required to move to the next state. Moreover, note that in every iteration, CRNN updates its internal state $\mathcal{H}$ given an input which is discussed shortly. As such, CRNN also allows information to be propagated efficiently across iterations, in contrast to the sequential models using CNNs which collapse the intermediate feature representation to $\mathbf{z}^{(i)}$.

In order to exploit the dynamic nature and the temporal redundancy of our data, we further propose to jointly model the recurrence evolving over time for dynamic MRI reconstruction. The proposed CRNN-MRI network and CRNN block are shown in Fig. 5.2(a), in which CRNN block comprised of 5 components:

1. bidirectional convolutional recurrent units evolving over time and iterations (BCRNN-t-i),
2. convolutional recurrent units evolving over iterations only (CRNN-i),
3. 2D convolutional neural network (CNN),
4. residual connection and
5. DC layers.

We introduce details of the components of our network in the following subsections.

CRNN-i

As aforementioned, we encapsulate the iterative optimisation procedures explicitly with RNNs. In the CRNN-i unit, the iteration step is viewed as the sequential step in the vanilla RNN. If the network is unfolded over the iteration dimension, the network can be illustrated as in Fig. 5.2(b), where information is propagated between iterations. Here we use $\mathbf{H}$ to denote the feature representation of our sequence of frames throughout the network. $\mathbf{H}_l^{(i)}$ denotes the representation at layer l (subscript) and iteration step i (superscript). Therefore, at iteration (i), given the input $\mathbf{H}_{l-1}^{(i)}$ and the previous iteration's hidden state $\mathbf{H}_l^{(i-1)}$, the hidden state $\mathbf{H}_l^{(i)}$

at layer l of a CRNN-i unit can be formulated as:

$$\mathbf{H}_l^{(i)} = \sigma(\mathbf{W}_l * \mathbf{H}_{l-1}^{(i)} + \mathbf{W}_i * \mathbf{H}_l^{(i-1)} + \mathbf{B}_l). \tag{5.7}$$

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

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

In this work, we use a vanilla RNN [Elm90] to model the recurrence due to its simplicity. Note this can be naturally generalised to other RNN units, such as long short-term memory (LSTM) and gated recurrent unit (GRU), which are considered to have better memory properties, although using these units would significantly increase computational complexity.

BCRNN-t-i

Dynamic MR images exhibit high temporal redundancy, which is often exploited as a-priori knowledge to regularise the reconstruction. Hence, it is also beneficial for the network to learn the dynamics of sequences. To this extent, we propose a bidirectional convolutional recurrent unit (BCRNN-t-i) to exploit both temporal *and* iteration dependencies jointly. BCRNN-t-i includes three convolution layers: one on the input which comes into the unit from the previous layer indicated by the green arrows in Fig. 5.2(c), one on the hidden state from the past and future time frames as shown by the red arrows, and the one on the hidden state from the previous iteration of the unit (blue arrows in Fig. 5.2(c)). Note that we simultaneously consider temporal dependencies from past and future time frames, and the encoding weights are shared for both directions. The output for the BCRNN-t-i layer is obtained by summing the feature maps learned from both directions. The illustration figure of the unit when it is unfolded over time sequence is shown in Fig. 5.2(c).

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

$$\begin{aligned}
\overrightarrow{\mathbf{H}}_{l,t}^{(i)} &= \sigma(\mathbf{W}_l * \mathbf{H}_{l-1,t}^{(i)} + \mathbf{W}_t * \overrightarrow{\mathbf{H}}_{l,t-1}^{i} + \mathbf{W}_i * \mathbf{H}_{l,t}^{(i-1)} + \overrightarrow{\mathbf{B}}_l), \\
\overleftarrow{\mathbf{H}}_{l,t}^{(i)} &= \sigma(\mathbf{W}_l * \mathbf{H}_{l-1,t}^{(i)} + \mathbf{W}_t * \overleftarrow{\mathbf{H}}_{l,t+1}^{(i)} + \mathbf{W}_i * \mathbf{H}_{l,t}^{(i-1)} + \overleftarrow{\mathbf{B}}_l), \\
\mathbf{H}_{l,t}^{(i)} &= \overrightarrow{\mathbf{H}}_{l,t}^{(i)} + \overleftarrow{\mathbf{H}}_{l,t}^{(i)},
\end{aligned} \tag{5.8}$$

Similar to the notation in Section 5.4, $\mathbf{W}_t$ represents the filters of recurrent convolutions

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

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

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

5.4.1 Network learning

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

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n_S} \sum_{(\mathbf{x}_u, \mathbf{x}_t) \in S} \|\mathbf{x}_t - \mathbf{x}_{rec}\|_2^2 \tag{5.9}$$

where $\boldsymbol{\theta} = \{\mathbf{W}_l, \mathbf{W}_i, \mathbf{W}_t, \mathbf{B}_l\}$, $l = 1 \dots L$, and n_S stands for the number of samples in the training set S. Note that the total number of time sequences T and iteration steps N assumed by the network before performing the reconstruction is a free parameter that must be specified in advance. The network weights were initialised using He initialisation [He+15] and it was trained using the Adam optimiser [KB15]. During training, gradients were hard-clipped to the range of $[-5, 5]$ to mitigate the gradient explosion problem. The network was implemented in Python using Theano and Lasagne libraries.

5.5 Experiments

5.5.1 Dataset and implementation details

The proposed method was evaluated using a complex-valued MR dataset consisting of 10 fully sampled short-axis cardiac cine MR scans. Each scan contains a single slice SSFP acquisition with 30 temporal frames, which have a 320×320 mm field of view and 10 mm thickness. The raw data consists of 32-channel data with sampling matrix size 192×190, which was then zero-filled to the matrix size 256×256. The raw multi-coil data was reconstructed using SENSE [Pru+99] with no undersampling and retrospective gating. Coil sensitivity maps were normalised to a body coil image and used to produce a single complex-valued reconstructed image. In experiments, the complex valued images were back-transformed to regenerate k-space samples, simulating a fully sampled single-coil acquisition. The input undersampled image sequences were generated by randomly undersampling the k-space samples using Cartesian undersampling masks, with undersampling patterns adopted from [JYK07]: for each frame the eight lowest spatial frequencies were acquired, and the sampling probability of k-space lines along the phase-encoding direction was determined by a zero-mean Gaussian distribution. Note that the undersampling rates are stated with respect to the matrix size of raw data, which is 192×190.

The architecture of the proposed network used in the experiment is shown in Fig. 5.2: each

iteration of the CRNN block contains five units: one layer of BCRNN-t-i, followed by three layers of CRNN-i units, and followed by a CNN unit. For all CRNN-i and BCRNN-t-i units, we used a kernel size $k = 3$ and the number of filters was set to $n_f = 64$ for Proposed-A and $n_f = 128$ for Proposed-B in Table 5.1. The CNN after the CRNN-i units contains one convolution layer with $k = 3$ and $n_f = 2$, which projects the extracted representation back to the image domain which contains complex-valued images expressed using two channels. For all convolutional layers, we used stride = 1 and paddings with half the filter size (rounded down) on both size. The output of the CRNN block is connected to the residual connection, which sums the output of the block with its input. Finally, we used DC layers on top of the CRNN output layers. During training, the iteration step is set to be $N = 10$, and the time sequence for training is $T = 30$. Note that this architecture is by no means optimal and more layers can be added to increase the ability of our network to better capture the data structures (see Section 5.5.4 for comparisons).

The evaluation was done via a 3-fold cross validation, where for two folds we train on 7 subjects then test on 3 subjects, and for the remaining fold we train on 6 subjects and test on 4 subjects. While the original sequence has size $256 \times 256 \times T$, For the training, we extract patches of size $256 \times D_{patch} \times T$, where $D_{patch} = 32$ is the patch size and the direction of patch extraction corresponds to the frequency-encoding direction. Note that since we only consider Cartesian undersampling, the aliasing occurs only along the phase encoding direction, so patch extraction does not alter the aliasing artefact. Patch extraction as well as data augmentation was performed on-the-fly, with random affine and elastic transformations on the image data. Undersampling masks were also generated randomly following patterns in [JYK07] for each input. During test time, the network trained on patches is directly applied on the whole sequence of the original image. The minibatch size during the training was set to 1, and we observed that the performance can reach a plateau within 6×10^4 backpropagations.

5.5.2 Evaluation method

We compared the proposed method with the representative algorithms of the CS-based dynamic MRI reconstruction, such as k-t FOCUSS [JYK07] and k-t SLR [Lin+11], and two variants of 3D CNN networks named 3D CNN-S and 3D CNN in our experiments. The built baseline 3D CNN networks share the same architecture with the proposed CRNN-MRI network but all the recurrent units and 2D CNN units were replaced with 3D convolutional units, that is, in each iteration, the 3D CNN block contain 5 layers of 3D convolutions, one DC layer and a residual connection. Here 3D CNN-S refers to network sharing weights across iterations, however, this does not employ the hidden-to-hidden connection as in the CRNN-i unit. The 3D CNN-S architecture was chosen so as to make a fair comparison with the proposed model using a comparable number of network parameters. In contrast, 3D CNN refers to the network without weight sharing, in which the network capacity is $N = 10$ times of that of 3D CNN-S, and approximately 12 times more than that of our first proposed method (Proposed-A). For the 3D CNN approaches, the receptive field size is $11 \times 11 \times 11$, as the receptive field size is "reset" after each data consistency layer. In contrast, for the proposed method, due to the hidden connections between iterations and bidirectional temporal connections, by tracing the longest path of the convolution layers involved in the forward pass, including both temporal and iterative directions, in theory, the receptive field size is $309 \times 309 \times 30$ (154 layers of CNNs for the middle frame in a sequence of 30 frames). However, the network still may predominantly rely on local features coming from the partial reconstruction. Nevertheless, the RNN has the ability to exploit the features with larger filter size if needed, which is not the case for 3D CNNs.

Reconstruction results were evaluated based on the following quantitative metrics: MSE, peak-to-noise-ratio (PSNR), structural similarity index (SSIM) [Wan+04] and high frequency error norm (HFEN) [RB11]. The choice of the these metrics was made to evaluate the reconstruction results with complimentary emphasis. MSE and PSNR were chosen to evaluate the overall accuracy of the reconstruction quality. SSIM put emphasis on image quality perception. HFEN was used to quantify the quality of the fine features and edges in the reconstructions, and here

Table 5.1: Performance comparisons (MSE, PSNR:dB, SSIM, and HFEN) on dynamic cardiac data with different acceleration rates. MSE is scaled to 10^{-3}. The bold numbers are better results of the proposed methods than that of the other methods.

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

we employed the same filter specification as in [RB11; Mia+16] with the filter kernel size 15×15 pixels and a standard deviation of 1.5 pixels. For PSNR and SSIM, it is the higher the better, while for MSE and HFEN, it is the lower the better.

5.5.3 Results

The comparison results of all methods are reported in Table 5.1, where we evaluated the quantitative metrics, network capacity and reconstruction time. Numbers shown in Table 5.1 are mean values of corresponding metrics with standard deviation of different subjects in parenthesis. Bold numbers in Table 5.1 indicate the better performance of the proposed methods than the competing ones. Compared with the baseline method (k-t FOCUSS and k-t SLR), the proposed methods outperform them by a considerable margin at different acceleration rates. When compared with deep learning methods, note that the network capacity of Proposed-A is comparable with that of 3D CNN-S and the capacity of Proposed-B is around one third of that of 3D CNN. Though their capacities are much smaller, both Proposed-A and Proposed-B outperform 3D CNN-S and 3D CNN for all acceleration rates by a large margin, which shows the competitiveness and effectiveness of our method. In addition, we can see a substantial improvement of the reconstruction results on all acceleration rates and in all metrics when

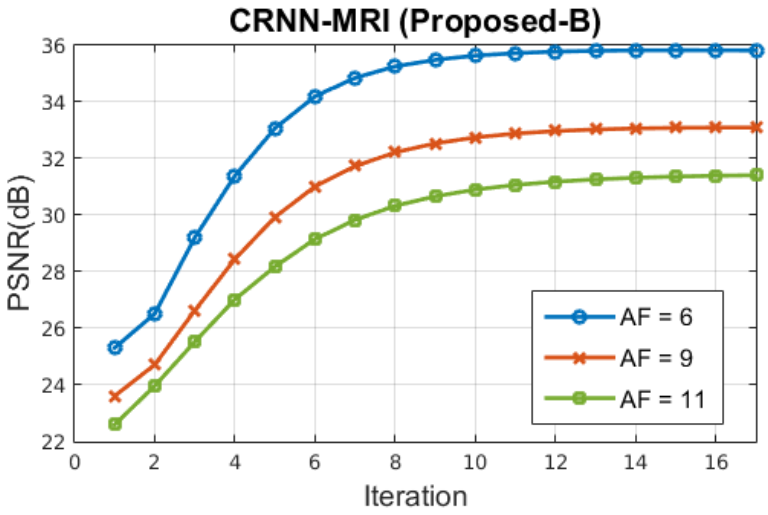

Figure 5.3: Mean PSNR values (Proposed-B) vary with the number of iterations at test time on data with different acceleration factors. Here AF stands for acceleration factor.

the number of network parameters is increased for the proposed method (Proposed-B), and therefore we will only show the results from Proposed-B in the following. The number of iterations used by the network at test time is set to be the same as the training stage, which is $N = 10$, however, if the iteration number is increased up to $N = 17$, it shows an improvement of 0.324dB on average. Fig. 5.3 shows the model's performance varying with the number of iterations at test time. Similarly, visualisation results of intermediate steps during the iterations of a reconstruction from 9× undersampling data are shown in Fig. 5.4, where we can observe the gradual improvement of the reconstruction quality from iteration step 1 to 10, which is consistent with the quantitative results as in Fig. 5.3.

A comparison of the visualisation results of a reconstruction from 9× acceleration is shown in Fig. 5.5 with the reconstructed images and their corresponding error maps from different reconstruction methods. As one can see, our proposed model (Proposed-B) can produce more faithful reconstructions for those parts of the image around the myocardium where there are large temporal changes. This is reflected by the fact that RNNs effectively use a larger receptive field to capture the characteristics of aliasing seen within the anatomy. Their temporal profiles at $x = 120$ are shown in Fig. 5.6. Similarly, one can see that the proposed model has overall much smaller error, faithfully modelling the dynamic data. It could be due to the fact that spatial and temporal features are learned separately in the proposed model while 3D CNN seeks

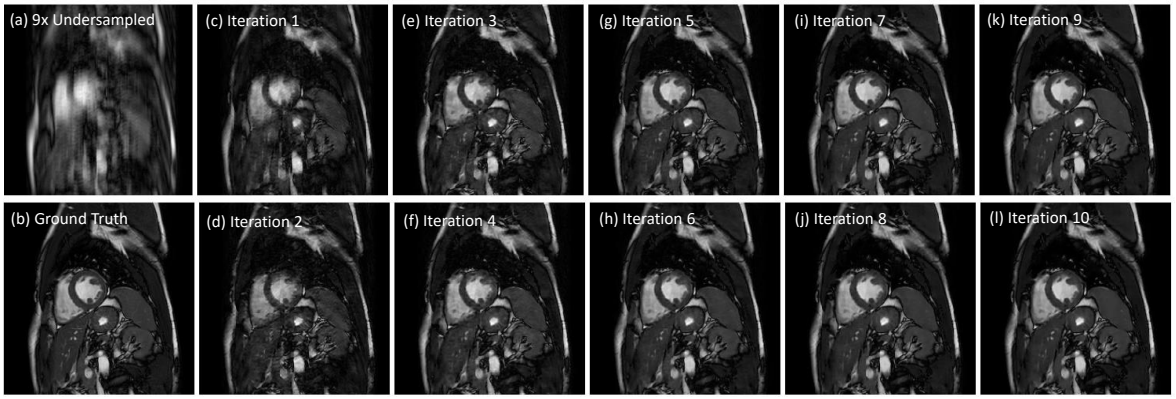

Figure 5.4: Visualisation results of intermediate steps during the iterations of a reconstruction. (a) Undersampled image by acceleration factor 9 (b) Ground Truth (c-l) Results from intermediate steps 1 to 10 in a reconstruction process.

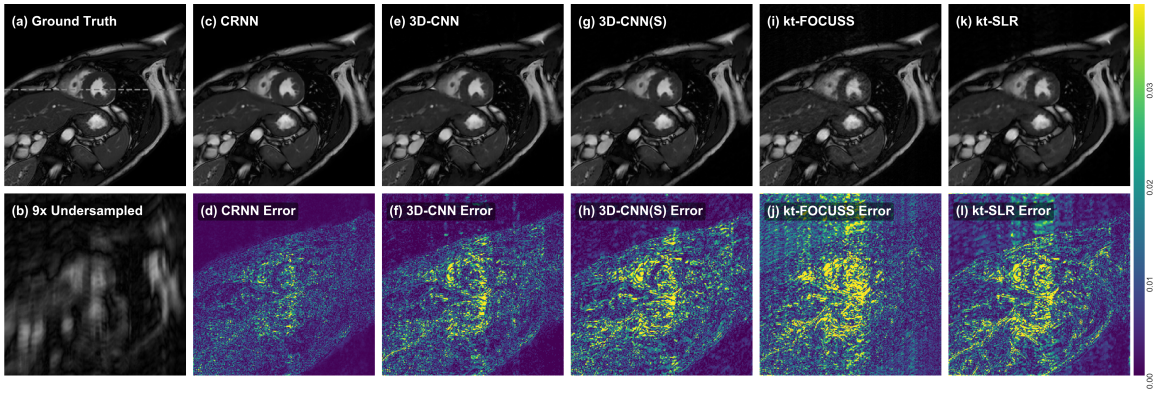

Figure 5.5: The comparison of reconstructions on spatial dimension with their error maps. (a) Ground Truth (b) Undersampled image by acceleration factor 9 (c,d) Proposed-B (e,f) 3D CNN (g,h) 3D CNN-S (i,j) k-t FOCUSS (k,l) k-t SLR

invariant feature learning across space and time.

In terms of speed, the proposed RNN-based reconstruction is faster than the 3D CNN approaches because it only performs convolution along time once per iteration, removing the redundant 3D convolutions which are computationally expensive. Reconstruction time of 3D CNN and the proposed methods reported in Table 5.1 were calculated on a GPU GeForce GTX 1080, and the time for k-t FOCUSS and k-t SLR were calculated on CPU.

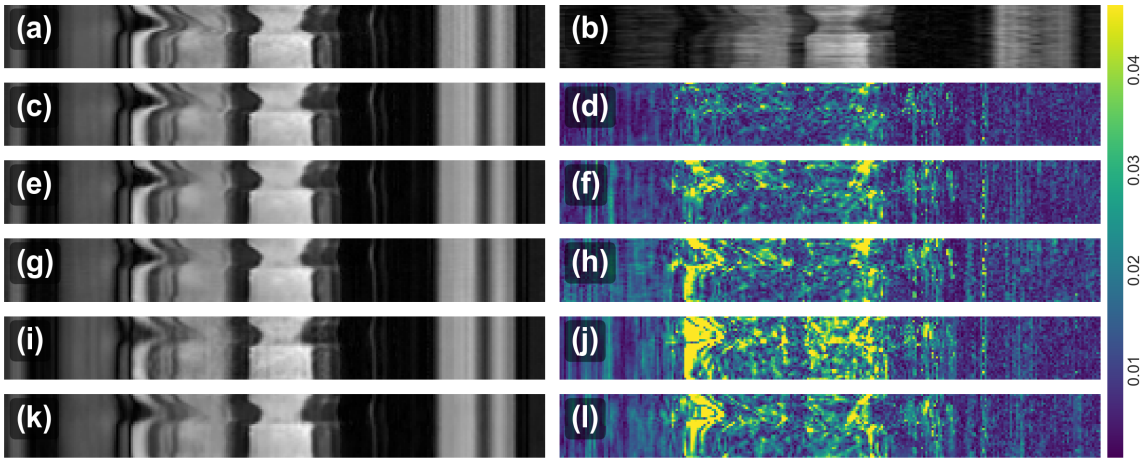

Figure 5.6: The comparison of reconstructions along temporal dimension with their error maps. (a) Ground Truth (b) Undersampled image by acceleration factor 9 (c,d) Proposed-B (e,f) 3D CNN (g,h) 3D CNN-S (i,j) k-t FOCUSS (k,l) k-t SLR

5.5.4 Variations of architecture

In this section we show additional experiments to investigate the variants of the proposed architecture. First, we study the effects of recurrence over iteration and time, separately and jointly. In this study, we performed experiments on data set with undersampling factor 9, and the number of iterations was set to be 2 in order to simplify and speed up the training. Results are shown in Table 5.2, where we present the mean PSNR value via 3-fold cross validation. To isolate the effects of both recurrence in the module, we proposed to remove one of the recurrence each time. By removing the recurrence over time, the network architecture degrades to 4 CRNN-i + CNN layers, and it doesn't exploit temporal information in this case. If the recurrence over iterations is removed, the network architecture then becomes BCRNN-t + 4 CNN layers, without any hidden connections between iterations. Note that in all architectures, the last CNN layer only has 2 filters, which is used to simply aggregate the latent representation back to image space. Therefore, we employ a simple convolution layer for this. From Table 5.2, it can be observed that by removing any of the recurrent connections, the performance becomes worse compared with the proposed architecture with both recurrence jointly. This indicate that both of these recurrence contribute to the learning of the reconstruction. In particular, it is also been observed that by removing the temporal recurrence, the network's

Table 5.2: Performance comparisons on investigating the effects of each recurrence in the module. Reported results are the mean PSNR on data with undersampling factor 9 via 3-fold cross-validation. For this study, the number of iteration was set as 2.

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

Table 5.3: Performance comparisons with different model architectures. Reported results are the mean PSNR on data with undersampling factor 9 via 3-fold cross-validation. (FPT: forward pass time; BPT: backward pass time)

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

performance degrades greatly compared with the one removing the iteration recurrence. This can be explained that by removing the temporal recurrence, the problem degrades to a single frame reconstruction, while dynamic reconstruction has been proven to be much better than single frame reconstruction as there exists great temporal redundancies that can be exploited between frames.

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

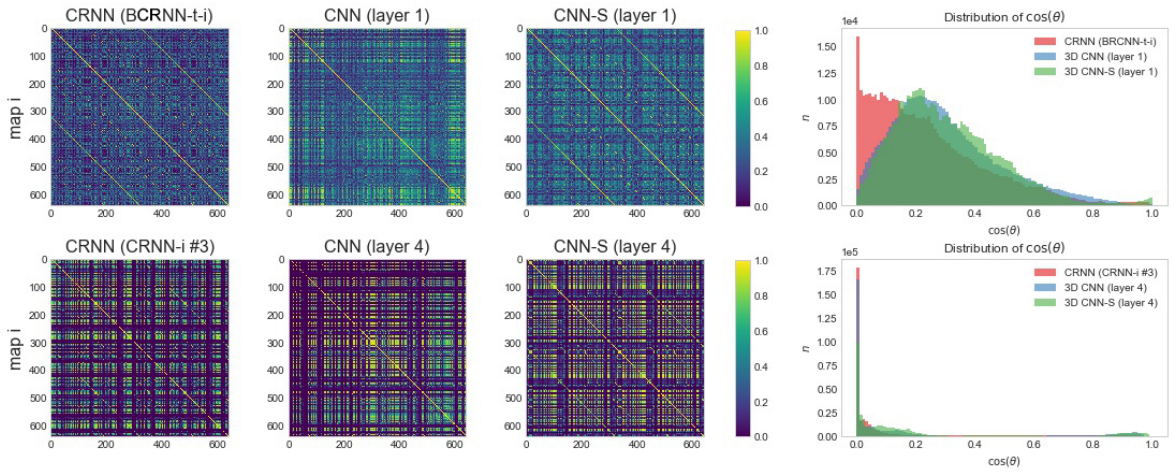

Figure 5.7: Cosine distances for the feature maps extracted from ith-layer of the subnetworks across 10 cascades/iterations. Top row shows $i = 1$, which corresponds to BCRNN-t-i unit for CRNN, 1st convolution layers for 3D-CNN and 3D-CNN-S. Bottom row shows $i = 4$, which corresponds to the third CRNN-i unit for CRNN, 4th convolution layers for 3D-CNN and 3D-CNN-S. In general, the distribution of $\cos(\theta)$ is closer to 0 for CRNN than for the CNN's.

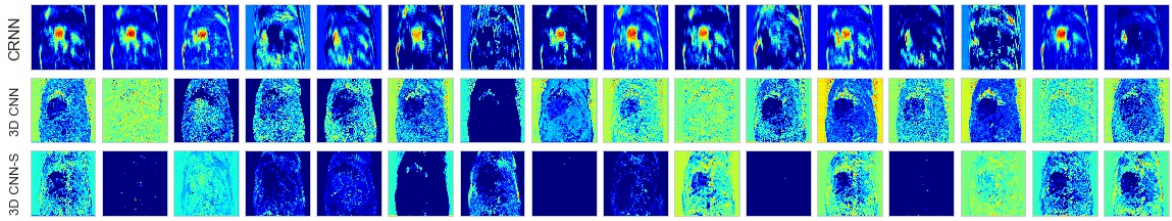

Figure 5.8: Examples of the feature maps from the CRNN-MRI (Proposed-A), 3D CNN and 3D CNN-S, at iteration 10

5.5.5 Feature map analysis

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

Geometrically, this supports the interpretation that if the cosine distance is small for all the feature map pairs, then the network is likely to be capturing diverse patterns. The result is summarised in Fig. 5.7, where the similarity measure is visualised as a matrix, as well as their distributions is plotted for each network.

We can see that for both $i \in \{1, 4\}$, the layers from CRNN appears to have geometrically more orthogonal feature maps. One can also observe that in general, layer 1 has higher redundancy compared to layer 4. In particular, the diagonal yellow stripes can be observed for CNN-S and CRNN, due to parameter-sharing for each cascade. This is not observed in 3D-CNN, even though many features do have high similarity. In Fig. 5.8 we show examples of the feature maps from layer 4 (3rd CRNN-i for CRNN, 4th convolution layers for 3D-CNN and 3D-CNN-S) at iteration/cascade 10 of each network during the forward pass. We selected 16 feature maps out of 64 by firstly clustering them into 16 groups, and then randomly chose one feature map from each group to show as representative feature maps in Fig. 5.8. These feature maps show the activations learned from different networks and are colour-coded (blue corresponds to low activation whereas red corresponds to high activation). We see that CRNN's features look significantly different from CNN. In particular, one can observe that some are activated by the dynamic region, and some are particularly sensitive to regions around the left and/or right ventricle.

5.6 Discussion

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

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

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

In this work, we modeled the recurrence using the relatively simple (vanilla) RNN architecture. For the future work, we will explore other recurrent units such as LSTM or GRU. As they are trained to explicitly select what to remember, they may allow the units to better control the flow of information and could reduce the number of iterations required for the network to generate high-quality output. Also, incorporating recurrent redundancy in k-space domain into

the proposed CRNN-MRI network is likely to improve the result, and will form part of our future work. In addition, we have found that the majority of errors between the reconstructed image and the fully sampled image lie at the part where motion exists, indicating that motion exhibits a challenge for such dynamic sequence reconstruction. Thus it will be interesting to explore more efficient ways that can improve the reconstruction quality while faithfully preserving cardiac motion. Additionally, current analysis only considers a single coil setup. In the future, we will also aim at investigating such methods in a scenario where multiple coil data from parallel MR imaging can be used jointly for higher acceleration acquisition.

5.7 Conclusion

Inspired by variable splitting and alternate minimisation strategies, we have presented an end-to-end deep learning solution, CRNN-MRI, for accelerated dynamic MRI reconstruction, with a forward, CRNN block implicitly learning iterative denoising interleaved by data consistency layers to enforce data fidelity. In particular, the CRNN architecture is composed of the proposed novel variants of convolutional recurrent unit which evolves over two dimensions: time and iterations. The proposed network is able to learn both the temporal dependency and the iterative reconstruction process effectively, and outperformed the other competing methods in terms of both reconstruction accuracy and speed for different undersampling rates.

Several key points were highlighted in this chapter. Firstly, dynamic data inherently carries the sufficient information needed for near-perfect reconstruction even from 10-fold accelerated data. Secondly, deep learning indeed can extract such information even from small number of parameters, as long as the network architecture enables that. This suggests that deep learning indeed has the capacity to directly infer information for subsequent analysis, provided that good network architecture can be found, which can extract such information. This is what is explored in the next chapter, where we perform segmentation of cardiac images directly from undersampled k-space data.

Chapter 6

Deep learning for direct cardiac segmentation from k-space data

This section is based on the following publications:

- **Schlemper, J.**, Oktay, O., Bai, W., Castro, D.C., Duan, J., Qin, C., Hajnal, J.V. and Rueckert, D., 2018, September. Cardiac MR segmentation from undersampled k-space using deep latent representation learning. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 259-267). Springer, 2018.

6.1 Introduction

Cardiovascular MR (CMR) imaging enables accurate quantification of cardiac chamber volume, ejection fraction and myocardial mass, which are crucial for diagnosing, assessing and monitoring cardiovascular diseases (CVDs), the leading cause of death globally. However, as aforementioned in the previous chapters, one limitation of CMR is the slow acquisition time. A routine CMR protocol can take from 20 to 60 minutes, which makes the tool costly and less accessible to worldwide population. In addition, CMR often requires breath-holds which can be difficult for patients; therefore, accelerating the CMR acquisition is essential. Over the

last decades, numerous approaches have been proposed for accelerated MR imaging, including parallel imaging, compressed sensing [Lus+08] and, more recently, deep learning approaches [Ham+18].

Reconstructing images from accelerated and undersampled MRI is an ill-posed problem and, essentially, all approaches must exploit some type of redundancies or assumptions on underlying data to resolve the aliasing caused by sub-Nyquist sampling. In the case of dynamic cardiac cine reconstruction, high spatiotemporal redundancy and sparsity can be exploited, however, the acceleration factor for a near perfect reconstruction is currently limited up to 9 [Sch+18a], as seen from the previous chapters. We argue that one effective way of pushing the acceleration factor even higher is to move to the concept of *application-driven MRI* [Cab+14a]. The key insight is that in many cases, the images are not an end in themselves, but rather means of accessing clinically relevant parameters which are obtained as post-processing steps, such as segmentation or tissue characterisation. Therefore, it is more effective to instead combine the reconstruction and post-processing steps and tailor the acquisition protocol to obtain the final results as accurately and efficiently as possible. In particular, if the end-goal is significantly more compressible than the original image, then one can expect further acceleration and still obtain satisfactory results [GG15; Guo+17]. This work focuses on a scenario where we obtain cardiac segmentation maps directly from heavily undersampled dynamic MR data.

Our contribution is the following: firstly, we propose two network architectures to learn such a mapping. The first model, *Syn-net*, exploits the spatiotemporal redundancy of the input to directly generate the segmentation map. However, under heavy aliasing artefact, the extracted features may not be useful for segmentation. To address the latter case, we propose the second model, *LI-net*, which first predicts the low dimensional latent code of the corresponding segmentation map, which is subsequently decoded. Secondly, we extensively evaluate the two models with large-scale simulation studies to demonstrate the effectiveness of the proposed approaches for various acceleration factors. In particular, we show that for the case where undersampled image contains sufficient geometrical information, Syn-net outperforms LI-net but in a more challenging scenario where only one line of k-space is sampled per frame, LI-net outperforms Syn-net. Finally, we study the latent space structure of these architectures to demonstrate that

the models learn useful representations of the data. This work potentially enables interesting future works in which reconstruction, post-processing and analysis stages are integrated to yield smarter imaging protocols.

6.2 Proposed methods

End-to-End Synthesis Network (Syn-net): Let $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$ be dataset of fully-sampled complex valued (dynamic) images $x = \{x_i \in \mathbb{C} \mid i \in S\}$, where S denotes indices on a pixel grid, and the corresponding segmentation labels $y = \{y_i \mid i \in S\}$ represent different tissue types with $y_i \in \{1, 2, \dots C\}$. Let $u = \{u_i \in \mathbb{C} \mid u = F_u^H F_u x\}$ denote an undersampled image, where F_u is the undersampling Fourier encoding matrix. Let $p(y_i \mid x)$ be the true distribution of i-th pixel label given an image, and $r(u \mid x, \mathcal{M})$ represent the sampling distribution of the undersampled images given an image x and a (pseudo) random undersampling mask generator $\mathcal{M}$. We aim to learn a synthesis network $q(y_i \mid u, \theta)$, termed *Syn-net*, which uses a convolutional neural network (CNN) to model the probability distribution of segmentation maps given the undersampled image parameterised by θ. We train the network by the following modified cross-entropy (CE) loss:

$$\mathcal{L}(\theta) = \sum_{(x,y)\in\mathcal{D}} \mathbb{E}_{u\sim r} \left[\sum_{i\in S} p(y_i \mid x_i) \log q(y_i \mid u_i, \theta) \right], \tag{6.1}$$

where we take the expectation over differently undersampled images. In practice, we generate one different undersampling pattern on-the-fly for each mini-batch training as an approximation to the expectation. For the network architecture, we use an architecture inspired by the state-of-the-art segmentation network, *U-net* [RFB15], shown in Fig. 6.1.

Latent Feature Interpolation Network (LI-net): Syn-net assumes that the input data contains sufficient geometrical information to generate the target segmentation. For heavily undersampled (and therefore aliased) images, this assumption may not be valid as the aliasing could mislead the network from identifying the correct boundaries. In the latter case, synthesis

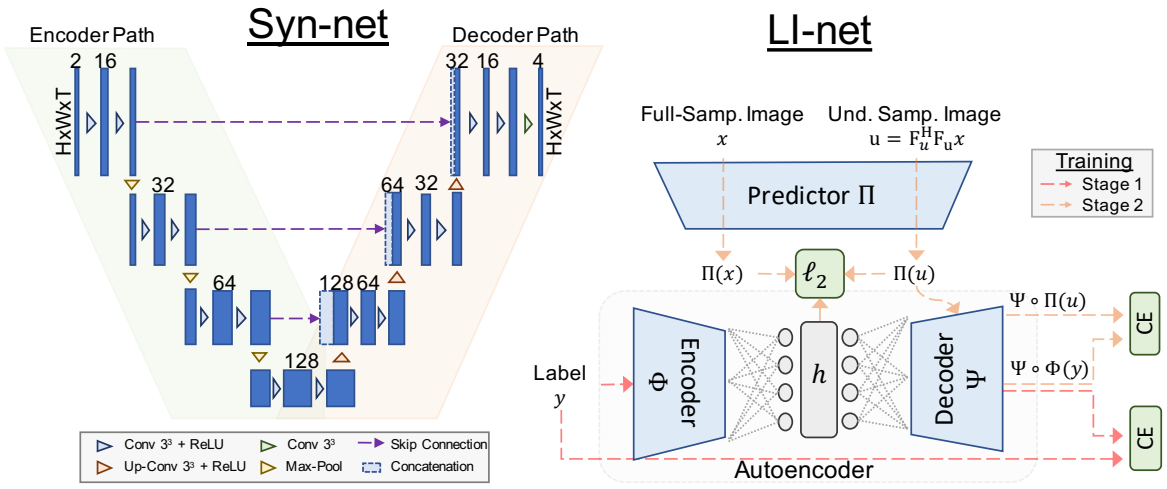

Figure 6.1: *(Left)* The detailed architecture of Syn-net: the changes in the number of features are shown above the tensor. *(Right)* For LI-net, the two-stage training strategy is outlined. The same encoder and decoder as Syn-net can be used for LI-Net

is still possible as long as the target domain has a compact, discriminative latent representation $h \in \mathcal{H}$ that can be predicted, an approach motivated by *TL-network* [Gir+16]. Such a network can be trained in following steps. In stage 1, one trains an auto-encoder (AE) in the target domain $y = \Psi(\Phi(y;\theta_{enc});\theta_{dec})$, $y \in \mathcal{Y}$, which is a composition of encoder $\Phi : \mathcal{Y} \to \mathcal{H}$ and decoder $\Psi : \mathcal{H} \to \mathcal{Y}$, parameterised by θ_{enc} and θ_{dec} respectively and $\mathcal{H}$ is a low-dimensional latent space. The AE can be trained using the ℓ_2 norm or CE loss. In stage 2, one trains a predictor network $\Pi : \mathcal{X} \to \mathcal{H}$, parameterised by θ_{pred}. For a given input-target pair (x, y), the predictor attempts to predict the latent code $h = \Phi(y;\theta_{enc})$ from x. This is trained using the ℓ_2 norm in the latent space: $d_{\mathcal{H}}(y, x) = \|\Phi(y;\theta_{enc}) - \Pi(x;\theta_{pred})\|_2$. Once the predictor is trained, one can obtain an input-output mapping by the composition $\hat{y} = \Psi(\Pi(x;\theta_{dec});\theta_{pred})$.

In our work, the AE is trained to learn the compact representation of segmentations and the predictor is trained to interpolate these from dynamic undersampled images, hence termed a *latent feature interpolation network* (*LI-net*). In stage 1, we train the AE using CE loss. In stage 2, we modify our objective to further encourage the network to produce a *consistent* prediction for differently undersampled versions of the same reference image. This constraint is implemented by forcing the network to produce the same latent code for undersampled images as for the fully-sampled image. Furthermore, we add a CE term $d_{\text{CE}}(y, \Psi \circ \Pi(u))$ to ensure that

an accurate segmentation can be obtained from the code. Therefore, our objective term is as follows (here λ_i's are hyper-parameters to be adjusted based on the preferred end-goal):

$$\mathcal{L}(\theta_{pred}) = \sum_{(x,y)\in\mathcal{D}} \mathbb{E}_{u\sim r}\Big[\, d_{\mathcal{H}}(y,u) + \lambda_1 d_{\mathcal{H}}(y,x) + \lambda_2 \|\Pi(x) - \Pi(u)\|_2 + \lambda_3 d_{\mathrm{CE}}(y, \Psi\circ\Pi(u))\Big]. \quad (6.2)$$

6.3 Experiments and results

Dataset and Undersampling: Experiments were performed using 5000 short-axis cardiac cine MR images from the UK Biobank study [Pet+16], which is acquired using bSSFP sequence, matrix size $N_x \times N_y \times T = 208 \times 187 \times 50$, a pixel resolution of $1.8{\times}1.8{\times}10.0\,\mathrm{mm}^3$ and a temporal resolution of 31.56 ms. Since the manual annotations are only available at end-systolic (ES) and end-diastolic (ED) frames but we are interested in segmenting the entire time sequence, we use [Bai+18], which well agrees with the manual segmentations, to generate the labels for the left-ventricular (LV) cavity, the myocardium and the right-ventricular (RV) cavity for all time-frames including apical, mid and basal slices, which were then treated as the ground truth labels for this work. We split the data into 4000 training subjects and 1000 test subjects, and we simulated random undersampling using variable density 1D undersampling masks. These masks were generated on-the-fly. As only the magnitude images were available we synthetically generated the phase maps (smoothly varying 2D sinusoid waves) on-the-fly to make the simulation more realistic by removing the conjugate symmetry in k-space. Different levels of acceleration factors ($1/n_l$) were considered, $n_l \in [1, 100]$ where n_l is the number of lines per time-frame. Note for fully-sampled image, $n_l = 168$.

Model and Parameters: The input to the network is 2D+t undersampled data and the output is a sequence of segmentation map. Note that z-slices were processed separately due to large slice thickness. The detail of the Syn-net is shown in Fig. 6.1. To make a fair comparison between the two architectures, we used the encoding path of Syn-net as both encoder Φ and

predictor Π, and the decoding path as decoder Ψ. The size of the latent code was set to be $|h| = 1024$. Note that fully-connected layers are used to join the encoder, the latent code and the decoder. They were trained with mini-batch size 8 using Adam with initial learning rate 10^{-4}, which was reduced by a factor of 0.8 every 2 epochs. The AE in LI-net was trained for 30 epochs to ensure that the Dice scores for each class reached 0.95. For both models, we first trained the network to perform segmentation from fully-sampled data as a warm start. The number of lines was gradually reduced and by 10^{th} epoch, we uniformly sampled n_l from $[0, 168]$. The training error for both models plateaued within 50 epochs. For LI-net, the hyper-parameters for the loss function were empirically chosen to be $\lambda_1 = 1$, $\lambda_2 = 10^{-4}$, $\lambda_3 = 10$, which we found to work sufficiently. For data augmentation, we generated affine transformations on-the-fly. We used PyTorch for implementation.

Evaluation: We first took the trained models and evaluated their Dice scores for LV, myocardium (Myo) and RV for $n_l \in [1, 100]$. For each subject, we only included ES and ED frames but aggregated the results across all short-axis slices. The Dice scores versus the number of acquired k-space lines are shown in Fig. 6.2. The networks maintained the performance up to about 20 lines per frame, demonstrating the capability of the models to directly interpolate the anatomical boundary even in the presence of the aliasing artefact. In general, Syn-net showed superior performance, indicating that the extracted spatiotemporal features are directly useful for segmentation. In particular, we report that the LI-net underperformed as it does not employ the skip-connection as Syn-net does, which limits how accurately it can delineate the boundaries. We speculate, however, increasing the capacity of network is likely to improve the results. Interestingly, LI-net outperformed Syn-net for the case of segmentation from 1 line, suggesting that in more challenging domains the approach of LI-net to interpolate the latent code is still a viable option.

In the second experiment, the models were further fine-tuned for a fixed number of lines for $n_l \in \{1, 10, 20\}$ separately. From the obtained segmentation maps, we computed LV ES/ED volumes (ESV/EDV) RV ESV/EDV, LV mass (LVM) and ejection fraction (EF). The mean percentage errors across all test subjects were reported in Table 6.1. Syn-net consistently performs better

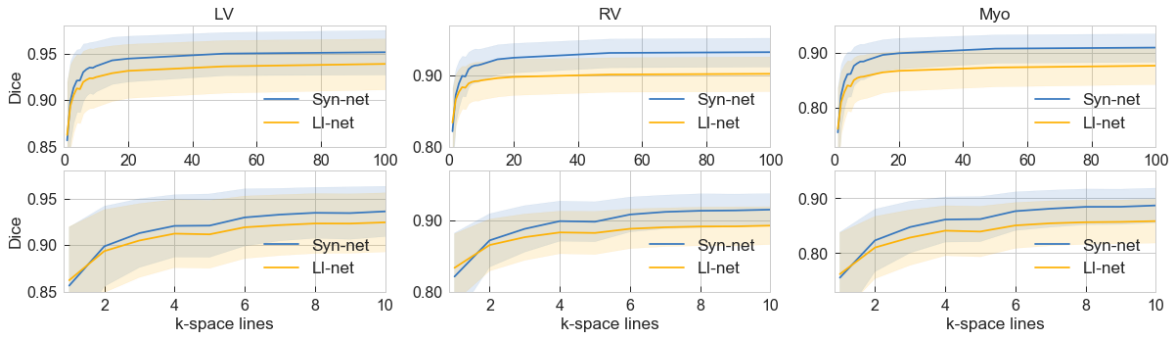

Figure 6.2: Dice scores of Syn-net vs LI-net. The second row expands $n_l \in [1, 10]$, the solid lines and the shaded areas show the mean and the standard deviation respectively.

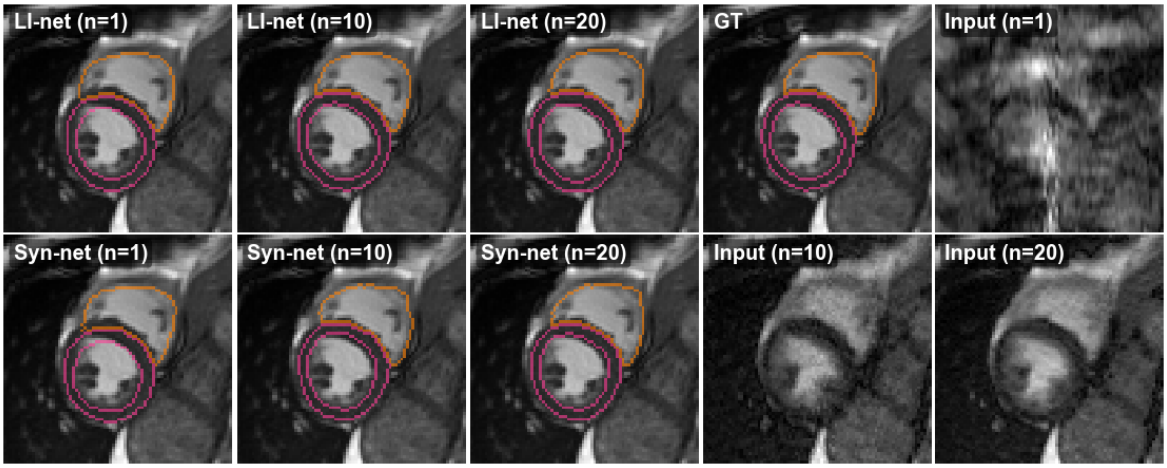

Figure 6.3: Visualisation of the ground truth image overlaid with the obtained segmentations. LI-net produced more anatomically regularised, consistent segmentations. Syn-net produced segmentations that are occasionally anatomically implausible but more faithful to the boundary.

than LI-net and has relatively small errors ($< 7.7\%$) for all values for $n_l \in \{10, 20\}$. Both models showed low error for EF, where the correlation coefficient was 0.81 for both models for $n_l = 20$. The examples of the segmentation maps are shown in Fig. 6.3. Note that due to heavy aliasing artefact of the input image, we instead visualised the temporally averaged image for x-y plane, which was obtained by combining all k-space lines across the temporal axis into a single k_x-k_y grid.

Although in theory we expect the reconstructed segmentation maps to be independent of the aliasing artefact present in the input, this is not always the case (Fig. 6.3). To measure such variability, we define *within subject distances*: given a fully-sampled image, we undersample it differently for $n_{\text{trial}} = 100$ times. From the predicted segmentation maps given by a model, we

Table 6.1: Average percentage errors (%) for each clinical parameter

	LV ESV			LV EDV			RV ESV			RV EDV			LVM			EF		
n_l	1	10	20	1	10	20	1	10	20	1	10	20	1	10	20	1	10	20
LI-net	7.9	3.6	3.2	15.8	7.9	7.0	11.7	6.5	6.6	18.4	10.5	10.6	25.0	13.7	12.9	8.8	5.5	4.9
Syn-net	9.0	4.2	3.4	14.6	7.2	6.1	9.9	4.9	4.1	13.4	7.7	6.5	11.4	6.8	5.8	8.2	5.5	4.6

Table 6.2: The *within-subject* and *between-subject* distances of the segmentations

	HD (Within)				MD (Within)				HD (Between)				MD (Between)			
	Myo		RV		Myo		RV		Myo		RV		Myo		RV	
n_l	1	10	1	10	1	10	1	10	1	10	1	10	1	10	1	10
LI-net	3.63	2.58	4.68	3.67	1.47	0.92	1.71	1.14	9.69	10.30	12.55	14.31	4.40	4.94	5.02	5.91
Syn-net	6.47	3.23	6.96	5.05	1.83	0.99	2.15	1.34	10.62	10.34	11.56	15.22	4.16	4.81	4.78	6.03

computed the mean shape, to which we then calculated mean contour distance (MD) and Hausdorff distance (HD) of individual predictions. Small distances indicate that the segmentation is consistent. However, if the network simply produces a *population mean shape* independent of the input, then the above distances can be very low even without producing useful segmentations. To get a better picture, we also measured the *between subject distances*, which computes MD and HD between the population mean shape (a mean predicted shape across *all* subjects) and the individual subject mean shapes. For both experiments, $n_{\text{subject}} = 100$ subject were used and the averaged distances are shown in Table 6.2. Indeed, we see that LI-net shows lower values for *within subject* distances, indicating that it produces more consistent segmentations than Syn-net ($p \ll 0.01$, Wilcoxon rank-sum). High *between subject* distances indicate that both models are generating segmentation maps closer to subject-specific means than to the population mean.

Finally, we investigate the latent space of the models. For 5 subjects, we generated 50 undersampled images for each number of lines $n_l \in \{1, 5, 10, 15, 20\}$. Here all undersampled images have the same target segmentation per subject. For LI-net, we plotted the predicted latent code $h \in \mathcal{H}$ for these images. For Syn-net, we plotted the activation map before the first upsampling layer to see whether the network exploits any latent space structure for generating the segmentations. We visualised them using Principal Component Analysis (PCA) and t-distributed

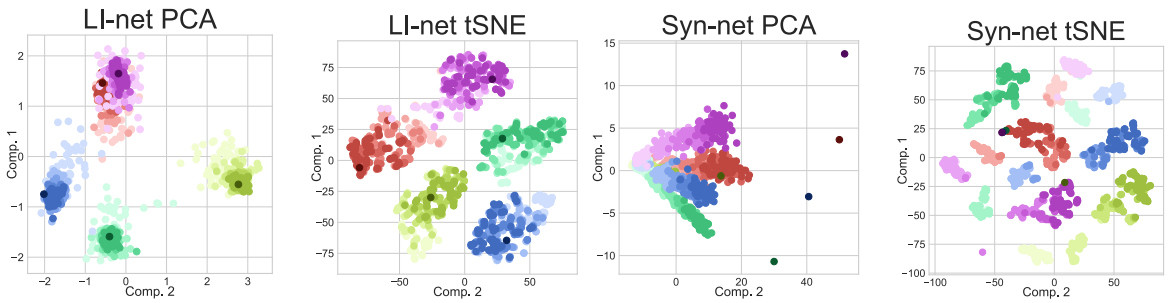

Figure 6.4: Visualising the distribution of the latent representations of LI-net and Syn-net. *(Left to right)* LI-net PCA, LI-net t-SNE, Syn-net PCA, Syn-net t-SNE. The darkest points are the latent representations of the fully sampled images, for reference.

stochastic neighbour embedding (t-SNE) with $d = 2$, as shown in Fig. 6.4, where subjects are colour-coded and brighter means higher acceleration factor.

For LI-net, both for PCA and t-SNE, the latent space is clearly clustered by individual subjects, indicating that the predictions are indeed consistent for different undersampling patterns. In addition, as the latent code is discriminative for each subject, it enables fitting a classifier for subject-based prediction tasks. On the other hand, for Syn-net, although there are per-subject clusters, there is also a clear tendency to favour clustering the points by the acceleration factors, as seen in the t-SNE plot. Note that since Syn-net also exploits skip connections, one can conclude that the network exploits different reconstruction strategies for different acceleration factors. Another interesting observation is that in Syn-net PCA, the distances between all the points are reduced as the acceleration factor is increased. This means that the latent features for Syn-net are less discriminative when images are heavily aliased. However, the extracted features become gradually more discriminative as more lines are sampled.

6.4 Conclusion and discussion

In this work we explored an application-driven MRI, where our end-goal was to extract segmentation maps directly from extremely undersampled data, bypassing image reconstruction. Remarkably, when at least 10 lines per frame are acquired, we showed that we could already compute clinical parameters within 10% error. Even though Syn-net provided better perfor-

mance overall, and LI-net exhibited more well-behaved latent-space structure. In future work, the latent code of LI-net could be used as a feature for classification tasks, where we may be able to classify whether a patient is abnormal, directly from a few lines of k-space. This work opens a huge avenue for future research where joint pipelines can be exploited for smarter MR imaging that is both fast and accurate.

This chapter highlighted that the network can indeed extract the quantitative values directly, however, the result depended on the network architecture. For example, LI-net managed to extract regular shapes, however, lost spatial information. On the other hand, Syn-net was sensitive to noise and aliasing. How can we characterise the behaviour of the network based on its architecture? How can one provide a consistent approach to understand what these models are doing, so we know when the network fails, why that was the case? Indeed, it is crucial to be able to explain the mechanics of these networks. The next chapter investigates one possibility of providing the explainability to the deep learning models using *attention-gates*.

Chapter 7

Attention models for interpretable automated methods

This section is based on the following publications:[1]

- **Schlemper, J.**[†], Oktay, O.[†], Chen, L., Matthew, J., Knight, C., Kainz, B., Glocker, B., Rueckert, D., Attention-gated networks for improving ultrasound scan plane detection., International Conference on Medical Imaging with Deep Learning, 2018.

- Oktay, O.[†], **Schlemper, J.**[†], Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., Rueckert, D., Attention U-Net: learning where to look for the pancreas. International Conference on Medical Imaging with Deep Learning, 2018.

- **Schlemper, J.**[†], Oktay, O.[†], Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis, 53, 197-207.

[1] [†] the authors contributed equally.

7.1 Introduction

Automated medical image analysis has been extensively studied in the medical imaging community due to the fact that manual labelling of large amounts of medical images is a tedious and error-prone task. Accurate and reliable solutions are required to increase clinical work flow efficiency and support decision making through fast and automatic extraction of quantitative measurements.

With the advent of convolutional neural networks (CNNs), near-radiologist level performance can be achieved in automated medical image analysis tasks including classification of Alzheimer's disease [Sar+17], skin lesions [Est+17; KH16] and echo-cardiogram views [Mad+18], lung nodule detection in CT/X-ray [Lia+17; Zhu+18b] and cardiac MR segmentation [Bai+18]. An extensive list of applications can be found in [Lit+17; Zah+18]. High representation power, fast inference, and weight sharing properties have made CNNs the de facto standard for image classification and segmentation.

Methods for existing applications rely heavily on multi-stage, cascaded CNNs when the target organs show large inter-patient variation in terms of shape and size. Cascaded frameworks extract a region of interest (ROI) and make dense predictions on that particular ROI. The application areas include cardiac MRI [KKK18], cardiac CT [Pay+17], abdominal CT [Rot+17; Rot+18] segmentation, and lung CT nodule detection [Lia+17]. However, this approach leads to excessive and redundant use of computational resources and model parameters; for instance, similar low-level features are repeatedly extracted by all models within the cascade.

To address this general problem, we propose a simple and yet effective solution, named *attention gate* (AG). By incorporating AGs into standard CNN models, model parameters and intermediate feature maps are expected to be utilised more efficiently while minimising the necessity of cascaded models to solve localisation and classification tasks separately. In more detail, AGs automatically learn to focus on target structures without additional supervision. At test time, these gates generate soft region proposals implicitly on-the-fly and highlight salient features useful for a specific task. In return, the proposed AGs improve model sensitivity and accuracy

for global and dense label predictions by suppressing feature activations in irrelevant regions. In this way, the necessity of using an external organ localisation module can be eliminated while maintaining the high prediction accuracy. In addition, they do not introduce significant computational overhead and do not require a large number of model parameters as in the case of multi-model frameworks. CNN models with AGs can be trained from scratch in a standard way similar to the training of fully convolutional network (FCN) models. Similar attention mechanisms have been proposed for natural image classification [Jet+18] and captioning [And+17] to perform adaptive feature pooling, where model predictions are conditioned only on a subset of selected image regions. In this work, we generalise this design and propose image-grid based gating that allows attention coefficients to be specific to local regions.

We demonstrate the performance of AG in real-time fetal ultrasound scan plane detection and CT pancreas segmentation. The first task is challenging due to low interpretability of the images and localising the object of interest is key to successful classification of the plane. To this end, we incorporate AGs into a variant of a VGG network, termed AG-Sononet, to demonstrate that attention mechanism can automatically localise the object of interest and improve the overall classification performance. The second task of pancreas segmentation is challenging due to low tissue contrast and large variability in organ shape and size. Moreover, we extend a standard U-Net architecture (*Attention U-Net*). We choose to evaluate our implementation on two commonly used benchmarks: TCIA Pancreas CT-82 [Rot+16] and multi-class abdominal CT-150. The results show that AGs consistently improve prediction accuracy across different datasets and training sizes while achieving state-of-the-art performance without requiring multiple CNN models.

7.1.1 Related work

Attention Gates: AGs are commonly used in classification tasks such as in the analysis of citation graphs [Vel+17] and natural images [Jet+18; Wan+17a]. Similarly in the context of natural language processing (NLP), such as image captioning [And+17] and machine translation [BCB14; LPM15; She+17; Vas+17], there have been several use cases of soft-attention models

to efficiently use the given context information. In particular, given a sequence of text and a current word, a task is to extract a next word in a sentence generation or translation. The idea of attention mechanisms is to generate a *context* vector which assigns weights on the input sequence. Thus, the signal highlights the salient feature of the sequence conditioned on the current word while suppressing the irrelevant counter-parts, making the prediction more contextualised.

Initial work on attention modelling has explored salient image regions by interpreting gradient of output class scores with respect to the input image. Trainable attention, on the other hand, is enforced by design and categorised as hard- and soft-attention. Hard attention [M+14], e.g. iterative region proposal and cropping, is often non-differentiable and relies on reinforcement learning for parameter updates, which makes model training more difficult. [YM17] used recursive hard-attention to detect anomalies in chest X-ray scans. Contrarily, soft attention is probabilistic, end-to-end differentiable, and utilises standard back-propagation without need for posterior sampling. For instance, additive soft attention is used in sentence-to-sentence translation [BCB14; She+17] and more recently applied to image classification [Jet+18; Wan+17a].

In computer vision, attention mechanisms are applied to a variety of problems, including image classification [Jet+18; Wan+17a; Zha+17d], segmentation [RZ16], action recognition [Liu+17; Pei+16; Wan+17c], image captioning [Lu+16; Xu+15], and visual question answering [NHK16; Yan+15]. [HSS17] used channel-wise attention to highlight important feature dimensions, which was the top-performer in the ILSVRC 2017 image classification challenge. Similarly, non-local self attention was used by [Wan+17c] to capture long range dependencies.

In the context of medical image analysis, attention models have been exploited for medical report generation [Zha+17c; Zha+17b] as well as joint image and text classification [Wan+18]. However, for standard medical image classification, despite often the information to be classified are extremely localised, only a handful of works use attention mechanisms [Gua+18; Pes+17]. In these methods, either bounding box labels are available to guide the attention, or local context is extracted by a hard-attention model (i.e. region proposal followed by hard-cropping).

2D Ultrasound Scan Plane Detection: Fetal ultrasound screening is an important diag-

nostic protocol to detect abnormal fetal development. During screening examination, multiple anatomically standardised [NHS15] scan planes are used to obtain biometric measurements as well as identifying abnormalities such as lesions. Ultrasound suffers from low signal-to-noise ratio and image artefacts. As such, diagnostic accuracy and reproducibility is limited and requires a high level of expert knowledge and training. In the past, several approaches were proposed [Che+15; Yaq+15], however, they are computationally expensive and cannot be deployed for the real-time application. More recently, [Bau+16] proposed a CNN architecture called *Sononet*. It achieves very good performance in real-time plane detection, retrospective frame retrieval (retrieving the most relevant frame) and weakly supervised object localisation. However, it suffers from low recall value in differentiating different planar views of the cardiac chambers, which requires the method to be able to exploit the subtle differences in the local structure and it makes the problem challenging.

Pancreas Segmentation in 3D-CT Images: Early work on pancreas segmentation from abdominal CT used statistical shape models [CSL16; SNS16] or multi-atlas techniques [Oda+17; Wol+13]. In particular, atlas approaches benefit from implicit shape constraints enforced by propagation of manual annotations. However, in public benchmarks such as the TCIA dataset [Rot+16], Dice similarity coefficients (DSC) for atlas-based frameworks are relatively low, ranging from 69.6% to 73.9% [Oda+17; Wol+13]. A classification based framework was proposed by [Zog+15] to remove the dependency of atlas to image registration. Recently, cascaded multi-stage CNN models [Rot+17; Rot+18; Zho+17] have been proposed to address the problem. Here, an initial coarse-level model (e.g. U-Net or Regression Forest) is used to obtain a ROI and then a cropped ROI is used for segmentation refinement by a second model. Similarly, combinations of 2D-FCN and recurrent neural network (RNN) models are utilised by [Cai+17] to exploit dependencies between adjacent axial slices. These approaches achieve state-of-the-art performance in the TCIA benchmark ($81.2\% - 82.4\%$ DSC). Without using a cascaded framework, the performance drops between 2.0 and 4.4 DSC points. Recently, [Yu+17] proposed an iterative two-stage model that recursively updates local and global predictions, and both models are trained end-to-end. Besides standard FCNs, dense connections [Gib+17] and sparse convolutions [HBO18; HO17] have been applied to the CT pancreas segmentation prob-

lem. Dense connections and sparse kernels reduce computational complexity by requiring less number of non-zero parameters.

7.1.2 Contributions

In this work, we propose a novel soft-attention gating module that can be utilised in CNN based standard image analysis models for dense label predictions. Additionally, we explore the benefit of AGs to medical image analysis, in particular, in the context of image classification and segmentation. The contributions of this work can be summarised as follows:

- We take the attention approach proposed by [Jet+18] a step further by proposing grid-based gating that allows attention gates to be more specific to local regions. This improves performance compared to gating based on a global feature vector. Moreover, our approach is not only limited to adaptive pooling [Jet+18] but can be also used for dense predictions as in segmentation networks.

- We propose one of the first use cases of soft-attention in a feed-forward CNN model applied to a medical imaging task that is end-to-end trainable. The proposed attention gates can replace hard-attention approaches used in image classification [YM17] and external organ localisation models in image segmentation frameworks [KKK18; Oda+17; Rot+17; Rot+18]. This also eliminates the need for any bounding box labels and backpropagation-based saliency map generation used by [Bau+16].

- For classification, we apply the proposed model to real-time fetal ultrasound scan plane detection and show its superior classification performance over the baseline approach. We show that attention maps can used for fast (weakly-supervised) object localisation, demonstrating that the attended features indeed correlate with the anatomy of interest.

- For segmentation, an extension to the standard U-Net model is proposed that provides increased sensitivity without the need of complicated heuristics, while not sacrificing specificity. We demonstrate that accuracy improvements when using U-Net are consistent across different imaging datasets and training sizes.

- We demonstrate that the proposed attention mechanism provides fine-scale attention maps that can be visualised, with minimal computational overhead, which helps with interpretability of predictions.

7.2 Methodology

7.2.1 Convolutional neural network

CNNs are now the state-of-the-art method for many tasks including classification , localisation and segmentation [Bai+18; Kam+17; Kam+18; Lee+15; Lit+17; LSD15; RFB15; Rot+17; Rot+18; XT15; Zah+18]. CNNs outperform traditional approaches in medical image analysis while being an order of magnitude faster than, e.g., graph-cut and multi-atlas segmentation techniques [Wol+13]. The success of CNNs is attributed to the fact that (I) domain specific image features are learnt using stochastic gradient descent (SGD) optimisation, (II) learnt kernels are shared across all pixels, and (III) image convolution operations exploit the structural information in medical images in an optimal fashion. However, it remains difficult to reduce false-positive predictions for small objects that show large shape variability. In such cases, in order to improve the accuracy, current frameworks [Gua+18; KKK18; Rot+17; Rot+18] rely on additional preceding object localisation models to simplify the task into separate localisation and subsequent classification/segmentation steps, or guide the localisation using weak labels [Pes+17]. Here, we demonstrate that the same objective can be achieved by integrating attention gates (AGs) in a standard CNN model. This does not require the training of multiple models and a large number of extra model parameters. In contrast to the localisation model in multi-stage CNNs, AGs progressively suppress feature responses in irrelevant background regions without the requirement to crop a ROI between networks.

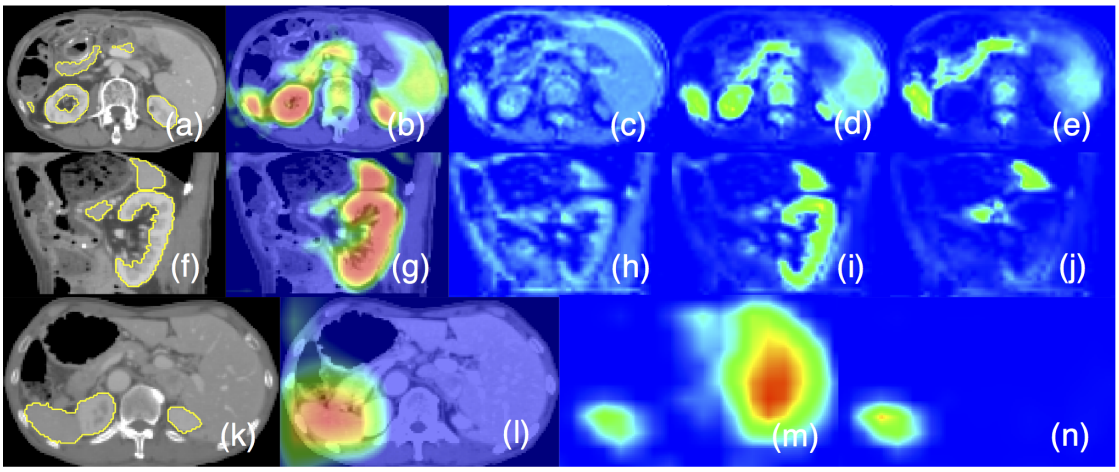

Figure 7.1: Axial (a) and sagittal (f) views of a 3DCT scan, (b,g) attention coefficients, image feature activations before (c,h) and after attention gating (d,e,i,j). Similarly, (k-n) visualise the gating on a coarse scale skip connection. The filtered feature activations (d,e,i,j) are collected from multiple AGs, where a subset of organs is selected by each gate and activations consistently correspond to specific structures across different scans.

7.2.2 Attention gate module

We now introduce *Attention Gate* (AG), which is a mechanism which can be incorporated in any existing CNN architecture. Let $\mathbf{x}^l = \{\mathbf{x}_i^l\}_{i=1}^n$ be the activation map of a chosen layer $l \in \{1, \dots, L\}$, where each $\mathbf{x}_i^l$ represents the pixel-wise feature vector of length F_l (i.e. the number of channels). For each $\mathbf{x}_i^l$, AG computes coefficients $\alpha^l = \{\alpha_i^l\}_{=1}^n$, where $\alpha_i^l \in [0, 1]$, in order to identify salient image regions and prune feature responses to preserve only the activations relevant to the specific task as shown in Figure 7.1. The output of AG is $\hat{\mathbf{x}}^l = \{\alpha_i^l \mathbf{x}_i^l\}_{i=1}^n$, where each feature vector is scaled by the corresponding attention coefficient.

The attention coefficients α_i^l are computed as follows: In standard CNN architectures, to capture a sufficiently large receptive field and thus, semantic contextual information, the feature-map is gradually downsampled. The features on the coarse spatial grid level identify location of the target objects and model their relationship at global scale. Let $\mathbf{g} \in \mathbb{R}^{F_g}$ be such global feature vector and provide information to AGs to disambiguate task-irrelevant feature content in $\mathbf{x}_i^l$. The idea is to consider each $\mathbf{x}_i^l$ and $\mathbf{g}$ jointly to attend the features at each scale l that are most relevant to the objective being minimised.

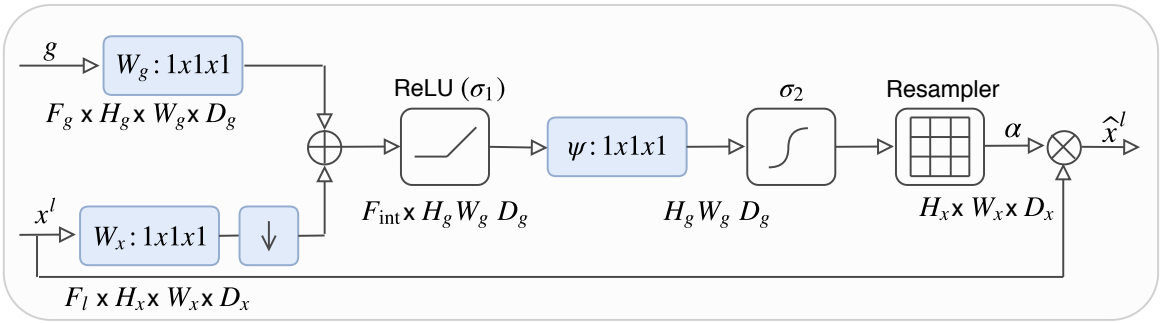

Figure 7.2: Schematic of the proposed additive attention gate (AG). Input features (x^l) are scaled with attention coefficients (α) computed in AG. Spatial regions are selected by analysing both the activations and contextual information provided by the gating signal (g) which is collected from a coarser scale. Grid resampling of attention coefficients is performed using trilinear interpolation.

There are two commonly used attention types: multiplicative [LPM15] and additive attention [BCB14]. The former is faster to compute and more memory-efficient in practice since it can be implemented as a matrix multiplication. However, additive attention is experimentally shown to be performing better for large dimensional input features [Bri+17]. For this reason, we use the latter to obtain the gating coefficient as can be seen in Figure 7.2, which is formulated as follows:

$$q^l_{att,i} = \boldsymbol{\psi}^T \left(\sigma_1 \left(\boldsymbol{W}_x^T \boldsymbol{x}^l_i + \boldsymbol{W}_g^T \boldsymbol{g} + \boldsymbol{b}_{xg} \right) \right) + b_\psi \tag{7.1}$$

$$\alpha^l = \sigma_2(q^l_{att}(\boldsymbol{x}^l , \boldsymbol{g} ; \boldsymbol{\Theta}_{att})), \tag{7.2}$$

where $\sigma_1(x)$ is an element-wise nonlinearity (e.g. rectified linear-unit) and $\sigma_2(x)$ is a normalisation function. For example, one can apply sigmoid to restrict the range to $[0, 1]$, or one can apply softmax operation $\alpha^l_i = e^{q^l_{att,i}} / \sum_i e^{q^l_{att,i}}$ such that the attention map sums to 1. AG is therefore characterised by a set of parameters $\boldsymbol{\Theta}_{att}$ containing: linear transformations $\boldsymbol{W}_x \in \mathbb{R}^{F_l \times F_{int}}$, $\boldsymbol{W}_g \in \mathbb{R}^{F_g \times F_{int}}$, $\boldsymbol{\psi} \in \mathbb{R}^{F_{int} \times 1}$ and bias terms $b_\psi \in \mathbb{R}$, $\boldsymbol{b}_{xg} \in \mathbb{R}^{F_{int}}$. The linear transformations are computed using channel-wise $1 \times 1 \times 1$ convolutions.

We note that AG parameters can be trained with the standard back-propagation updates without a need for sampling based optimisation methods as used in hard-attention [M+14]. While AG does not require auxiliary loss function to optimise, we found that using deep-supervision

[Lee+15] encourages the intermediate feature-maps to be semantically discriminative at each image scale. This ensures that attention units, at different scales, have an ability to influence the responses to a large range of image foreground content. We therefore prevent dense predictions from being reconstructed from small subsets of gated feature-maps.

Multi-dimensional attention

In case of where multiple semantic classes are present in the image, one can learn multi-dimensional attention coefficients. This is inspired by the approach of [She+17], where multi-dimensional attention coefficients are used to learn sentence embeddings. Thus, each AG learns to focus on a subset of target structures. In case of multi-dimensional AGs, each α^l corresponds to a vector and produce $\hat{\boldsymbol{x}}^l = [\alpha^l_{(1)} \odot \boldsymbol{x}^l, \ldots, \alpha^l_{(m)} \odot \boldsymbol{x}^l]$ where $\alpha^l_{(k)}$ is k-th sub AG and $\odot$ is element-wise multiplication operation. In each sub-AG, complementary information is extracted and fused to define the output of skip connection.

Gating signal and grid attention

As the gating signal $\mathbf{g}$ must encode global information from large spatial context, it is usually obtained from the coarsest scale activation map. For example in classification, one could use the activation map just before the final softmax layer. In the context of medical imaging, however, since most objects of interest are highly localised, flattening may have the disadvantage of losing important spatial context. In fact, in many cases a few max-pooling operations are sufficient to infer the global context without explicitly using the global pooling. Therefore, we propose a *grid attention* mechanism. The idea is to use the coarse scale feature map before any flattening is done. For example, given an input tensor size of $F_l \times H_x \times W_x$, after r max pooling operations, the tensor size is reduced to $F_g \times H_g \times W_g = F_g \times H_x/(2^r) \times W_y/(2^r)$. To generate the attention map, we can either downsample or upsample the coarse grid to match the spatial resolution of $\boldsymbol{x}^l$. In this way, the attention mechanism has more flexibility in terms of what to focus on a regional basis. For upsampling, we chose to use bilinear upsampling. Note that the upsampling can be replaced by a learnable weight, however, we did not opt for this for the sake

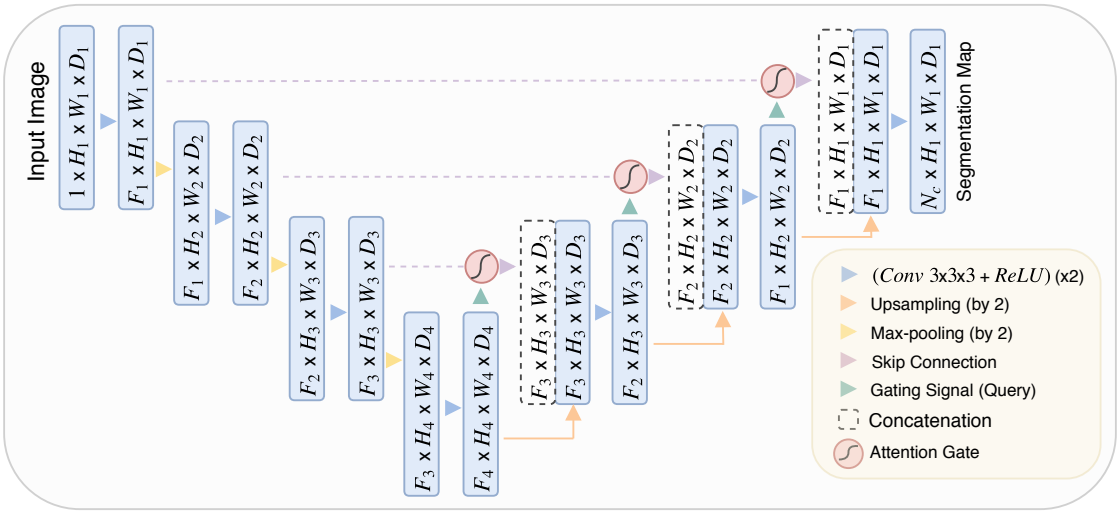

Figure 7.3: A block diagram of the proposed Attention U-Net segmentation model. Input image is progressively filtered and downsampled by factor of 2 at each scale in the encoding part of the network (e.g. $H_4 = H_1/8$). N_c denotes the number of classes. Attention gates (AGs) filter the features propagated through the skip connections. Schematic of the AGs is shown in Figure 7.2. Feature selectivity in AGs is achieved by use of contextual information (gating) extracted in coarser scales.

of simplicity. For segmentation, one can directly use the coarsest activation map as the gating signal.

Backward pass through attention gates

Information extracted from coarse scale is used in gating to disambiguate irrelevant and noisy responses in input feature-maps. For instance, in the U-Net architecture, gating is performed on skip connections right before the concatenation to merge only relevant activations. Additionally, AGs filter the neuron activations during the forward pass as well as during the backward pass. Gradients originating from background regions are down weighted during the backward pass. This allows model parameters in shallower layers to be updated mostly based on spatial regions that are relevant to a given task. The update rule for convolution parameters in layer $l-1$ can be formulated as follows:

$$\frac{\partial(\hat{x}_i^l)}{\partial\left(\Phi^{l-1}\right)} = \frac{\partial\left(\alpha_i^l\, f(x_i^{l-1};\Phi^{l-1})\right)}{\partial\left(\Phi^{l-1}\right)} = \alpha_i^l \frac{\partial(f(x_i^{l-1};\Phi^{l-1}))}{\partial\left(\Phi^{l-1}\right)} + \frac{\partial(\alpha_i^l)}{\partial\left(\Phi^{l-1}\right)} x_i^l \tag{7.3}$$

where the first gradient term on the right-hand side is scaled with α_i^l.

7.2.3 Attention gates for segmentation

In this work, we build our attention-gated segmentation model on top of a standard 3D U-Net architecture. U-Nets are commonly used for image segmentation tasks because of their good performance and efficient use of GPU memory. The latter advantage is mainly linked to extraction of image features at multiple image scales. Coarse feature-maps capture contextual information and highlight the category and location of foreground objects. Feature-maps extracted at multiple scales are later merged through skip connections to combine coarse- and fine-level dense predictions as shown in Figure 7.3. The proposed AGs are incorporated into the standard U-Net architecture to highlight salient features that are passed through the skip connections. For AGs, we chose sigmoid activation function for normalisation: $\sigma_2(x) = \frac{1}{1+exp(-x)}$. While in image captioning [And+17] and classification [Jet+18] tasks, the softmax activation function is used to normalise the attention coefficients σ_2, however, sequential use of softmax yields sparser activations at the output. For dense prediction task, we empirically observed that sigmoid resulted in better training convergence for the AG parameters.

7.2.4 Attention gates for classification

For attention-gated classification model, we chose *Sononet* [Bau+16] to be our base architecture, which is a variant of VGG network [SZ14]. The difference is that Sononet can be decoupled into feature extraction module and adaptation module. In the adaptation module, the number of channels are first reduced to the number of target classes C. Subsequently, the spatial information is flattened via channel-wise global average pooling. Finally, a softmax operation is applied to the resulting vector and the entry with maximum activation is selected as the prediction. As the network is constrained to classify based on the reduced vector, the network is forced to extract the most salient features for each class.

The proposed attention mechanism is incorporated in the Sononet architecture to better ex-

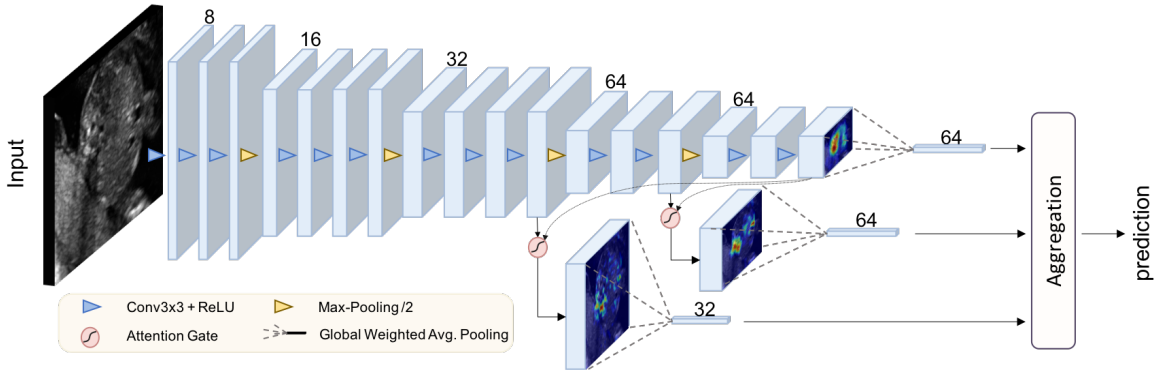

Figure 7.4: The schematics of the proposed attention-gated classification model, *AG-Sononet*. The proposed attention units are incorporated in layer 11 and layer 14. The attention maps are summed along the spatial axes, resulting in vectors with F_{l_i} features. The vectors are combined using fully connected layers at aggregation stage to yield final predictions.

ploit local information. In the modified architecture, termed Attention-Gated Sononet (*AG-Sononet*), we remove the adaptation module. The final layer of the feature extraction module is used as gridded global feature map $\mathbf{g}$. We apply the proposed attention mechanism to layer 11 and 14 just before pooling, We empirically found that attention gates were less effective if applied to the earliest layer. We speculate that this is because first few layers only represent low-level features, which is not discriminative yet to be attended. The proposed architecture is shown in Figure 7.4. After the attention coefficients $\{\alpha_i^l\}_{i=1}^n$ are obtained, the weighted average over the spatial axes is computed, yielding a vector of length F_l at scale l: $\tilde{\boldsymbol{x}}^l = \sum_{i=1}^n \alpha_i^l x_i^l$. In addition, we also perform the global average pooling on the coarsest scale representation. The prediction is given by fitting a fully connected layer on the concatenated feature vector $\{\tilde{\boldsymbol{x}}^{l_1}, \tilde{\boldsymbol{x}}^{l_2}, \tilde{\boldsymbol{x}}^{l_3}\}$ (e.g. $l_1 = 11, l_2 = 14, l_3 = 17$). We note that for AG-sononet, we normalised the attention coefficients as $\alpha_i^l = (\alpha_i^l - \alpha_{min}^l / \sum_j (\alpha_j^l - \alpha_{min}^l))$, where $\alpha_{min}^l = \min_j \alpha_j^l$, as we realised that softmax output was often too sparse, making the prediction more challenging.

Given the attended feature vectors at different scales, we highlight that the aggregation strategy is flexible and that it can be adjusted depending on the target problem. We empirically observed that a combination of deep-supervision [Lee+15] for each scale followed by fine-tuning using a new FC layer fitted on the concatenated vector gave the best performance.

The simplest is to just fit a fully connected layer on the concatenated vector as mentioned

above. However, we noticed that sometimes the network abandons the fine-scale attention mechanisms as it is non-trivial to train. An alternative approach is to fit a separate fully connected (FC) layer at each scale and make separate predictions. The final prediction is then given by either weighted mean or max operations. One can also use deep-supervision [Lee+15] to force each scale to learn a useful prediction as well as when combined. We empirically observed that first training the network at each scale, then fine-tuning using a new FC layer fitted on the concatenated vector worked the best. In the experimentation, we considered variations described below.

7.3 Experiments and results

The proposed AG model is modular and independent of application type; as such it can be easily adapted for pixel and image level classification tasks. To demonstrate its applicability to image classification and segmentation, we evaluate the proposed attention based FCN models on challenging abdominal CT multi-label segmentation and 2D ultrasound image plane classification problems. In particular, pancreas boundary delineation is a difficult task due to shape-variability and poor tissue contrast, similarly image quality and subject variability introduce challenges in 2D-US image classification. Our models are compared against the standard 3D U-Net and Sononet in terms of model prediction performance, model capacity, computation time, and memory requirements.

7.3.1 Evaluation datasets

In this section, we present the image datasets used in classification and segmentation experiments.

3D-CT abdominal image datasets

For the experiments, two different CT abdominal datasets are used: (I) 150 abdominal 3D CT scans acquired from patients diagnosed with gastric cancer (CT-150). In all images, the pancreas, liver, and spleen boundaries were semi-automatically delineated by three trained researchers and manually verified by a clinician. The same dataset is used by [Rot+17] to benchmark the U-Net model in pancreas segmentation. (II) The second dataset[2] (CT-82) consists of 82 contrast enhanced 3D CT scans with pancreas manual annotations performed slice-by-slice. This dataset (NIH-TCIA) [Rot+16] is publicly available and commonly used to benchmark CT pancreas segmentation frameworks. The images from both datasets are downsampled to isotropic 2.00 mm resolution due to the large image size and hardware memory limitations.

2D fetal ultrasound image dataset

Our dataset consisted of 2694 2D ultrasound examinations of volunteers with gestational ages between 18 and 22 weeks. The dataset contains 13 types of standard scan planes and background, complying the standard specified in the UK National Health Service (NHS) fetal anomaly screening programme (FASP) handbook [NHS15]. The standard scan planes are: Brain (Cb.), Brain (Tv.), Profile, Lips, Abdominal, Kidneys, Femur, Spine (Cor.), Spine (Sag.), 4CH, 3VV, RVOT, LVOT. The dataset further includes large portions of frames which contain anatomies that are not part of the scan plane, labelled as "background". The details of the image acquisition protocol as well as how scan plane labels are obtained can be found in [Bau+16]. The data was cropped to central 208×272 to prevent the network from learning the surrounding annotations shown in the ultrasound scan screen.

[2] `https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT`

7.3.2 Model training and implementation details

The datasets used in this manuscript contain large class imbalance issue that needs to be addressed. For ultrasound dataset, due to the nature of screening process, the background label dominates the dataset. To address this, we used a weighted sampling strategy, where we matched the probability of sampling one of the foreground labels to the probability of sampling a background label. For the segmentation models, the class imbalance problem is tackled using the Sorensen-Dice loss [Dro+16; MNA16] defined over all semantic classes. Dice loss is experimentally shown to be less sensitive to class imbalance in segmentation tasks.

For both tasks, batch-normalisation, deep-supervision [Lee+15], and standard data-augmentation techniques (affine transformations, axial flips, random crops) are used in training attention and baseline networks. Intensity values are linearly scaled to obtain a normal distribution $N(0, 1)$. For classification models, we empirically found that optimising with Stochastic Gradient Descent with Nesterov momentum ($\rho = 0.9$) worked the best. The initial learning rate was set to 0.1, which was subsequently reduced by a factor of 0.1 for every 100 epoch. We also used a warm-start learning rate of 0.01 for the first 5 epochs. For segmentation models, we used Adam with $\alpha = 10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$. The batch size for the Sononet models was set to 64. However, for the 3D-CT segmentation models, gradient updates are computed using small batch sizes of 2 to 4 samples. For larger segmentation networks, gradient averaging is used over multiple forward and backward passes. This is mainly because we propose a 3D-model to capture sufficient semantic context in contrast to the state-of-the-art CNN segmentation frameworks [Cai+17; Rot+18]. Gating parameters are initialised so that attention gates pass through feature vectors at all spatial locations. Moreover, we do not require multiple training stages as in hard-attention based approaches therefore simplifying the training procedure.

Implementation details:

The architecture for AG-sononet is shown in Fig. 7.4. The parameters for AG-Sononet was initialised using a partially trained Sononet. We compare our models with different capacities,

Table 7.1: Multi-class CT abdominal segmentation results obtained on the *CT*-150 dataset: The results are reported in terms of Dice score (DSC) and mesh surface to surface distances (S2S). These distances are reported only for the pancreas segmentations. The proposed Attention U-Net model is benchmarked against the standard U-Net model for different training and testing splits. Inference time (forward pass) of the models are computed for input tensor of size $160 \times 160 \times 96$. Statistically significant results are highlighted in bold font.

Method	U-Net	Att U-Net	U-Net	Att U-Net
Train/Test Split	120/30	120/30	30/120	30/120
Pancreas DSC	0.814±0.116	**0.840±0.087**	0.741±0.137	**0.767±0.132**
Pancreas Precision	0.848±0.110	0.849±0.098	0.789±0.176	**0.794±0.150**
Pancreas Recall	0.806±0.126	**0.841±0.092**	0.743±0.179	**0.762±0.145**
Pancreas S2S Dist (mm)	2.358±1.464	**1.920±1.284**	3.765±3.452	3.507±3.814
Spleen DSC	0.962±0.013	0.965±0.013	0.935±0.095	**0.943±0.092**
Kidney DSC	0.963±0.013	0.964±0.016	0.951±0.019	0.954±0.021
Number of Params	5.88 M	6.40 M	5.88 M	6.40 M
Inference Time	0.167 s	0.179 s	0.167 s	0.179 s

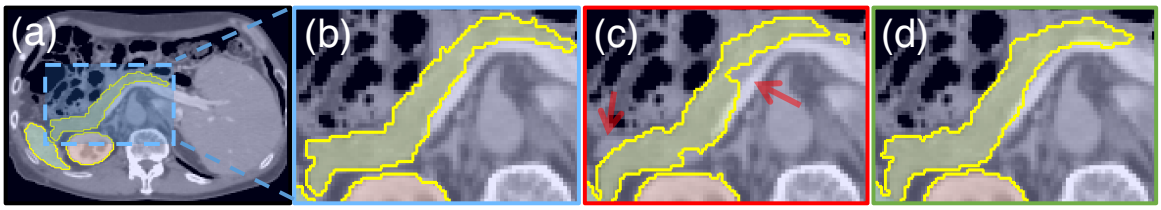

Figure 7.5: (a-b) The ground-truth pancreas segmentation, (c) U-Net and (d) Attention U-Net. The missed dense predictions by U-Net are highlighted with red arrows.

with the initial number of features 8, 16 and 32. For U-net and Attention U-net, the initial number of features is set to $F_1 = 8$, which is doubled after every max-pooling operation. Our implementation using PyTorch [Pas+17] is publicly available[3].

7.3.3 3D-CT abdominal image segmentation results

The proposed Attention U-Net model is benchmarked against the standard U-Net [RFB15] on multi-class abdominal CT segmentation. We use *CT*-150 dataset for both training (120) and testing (30). The corresponding Dice scores (DSC) and surface distances (S2S) are given in Table 7.1. The results on pancreas predictions demonstrate that attention gates (AGs) increase recall values ($p = .005$) by improving the model's expression power as it relies on

[3] `https://github.com/ozan-oktay/Attention-Gated-Networks`

Table 7.2: Segmentation experiments on CT-150 dataset are repeated with higher capacity U-Net models to demonstrate the efficiency of the attention models with similar or less network capacity. The additional filters in the U-Net model are distributed uniformly across all the layers. Segmentation results for the pancreas are reported in terms of dice score, precision, recall, surface distances. The models are trained with the same train/test data splits (120/30)

Method	# of Pars	DSC	Precision	Recall	S2S Dist (mm)	Run Time
U-Net	6.44 M	.821±.119	.849±.111	.814±.125	2.383±1.918	.191 s
U-Net	10.40 M	.825±.104	.861±.082	.807±.121	2.202±1.144	.222 s

Table 7.3: Pancreas segmentation results obtained on the TCIA Pancreas-CT Dataset [Rot+16]. The dataset contains in total 82 scans which are split into training (61) and testing (21) sets. The corresponding results are obtained before (BFT) and after fine tuning (AFT) and also training the models from scratch (SCR). Statistically significant results are highlighted in bold font.

	Method	Dice Score	Precision	Recall	S2S Dist (mm)
BFT	U-Net	0.690±0.132	0.680±0.109	0.733±0.190	6.389±3.900
	Attention U-Net	**0.712±0.110**	0.693±0.115	**0.751±0.149**	**5.251±2.551**
AFT	U-Net	0.820±0.043	0.824±0.070	0.828±0.064	2.464±0.529
	Attention U-Net	**0.831±0.038**	0.825±0.073	**0.840±0.053**	**2.305±0.568**
SCR	U-Net	0.815±0.068	0.815±0.105	0.826±0.062	2.576±1.180
	Attention U-Net	0.821±0.057	0.815±0.093	**0.835±0.057**	**2.333±0.856**

AGs to localise foreground pixels. The difference between predictions obtained with these two models are qualitatively compared in Figure 7.5. In the second experiment, the same models are trained with fewer training images (30) to show that the performance improvement is consistent and significant for different sizes of training data ($p = .01$). For both approaches, we observe a performance drop on spleen DSC as the training size is reduced. The drop is less significant with the proposed framework. For kidney segmentation, the models achieve similar accuracy since the tissue contrast is higher.

In Table 7.1, we also report the number of trainable parameters for both models. We observe that by adding 8% extra capacity to the standard U-Net, the performance can be improved by 2-3% in terms of DSC. For a fair comparison, we also train higher capacity U-Net models and compare against the proposed model with smaller network size. The results shown in Table 7.2 demonstrate that the addition of AGs contributes more than simply increasing model capacity (uniformly) across all layers of the network ($p = .007$). Therefore, additional capacity should be used for AGs to localise tissues, in cases when AGs are used to reduce the redundancy of

Table 7.4: State-of-the-art CT pancreas segmentation methods that are based on single and multiple CNN models. The listed segmentation frameworks are evaluated on the same public benchmark (CT-82) using different number of training and testing images. Similarly, the FCN approach proposed in [Rot+17] is benchmarked on CT-150 although it is trained on an external dataset (Ext).

Method	Dataset	Pancreas DSC	Train/Test	# Folds
Hierarchical 3D FCN [Rot+17]	CT-150	82.2 ± 10.2	Ext/150	-
Dense-Dilated FCN [Gib+17]	CT-82 & Synapse[4]	66.0 ± 10.0	63/9	5-CV
2D U-Net [HBO18]	CT-82	75.7 ± 9.0	66/16	5-CV
HN 2D FCN Stage-1[Rot+18]	CT-82	76.8 ± 11.1	62/20	4-CV
HN 2D FCN Stage-2[Rot+18]	CT-82	81.2 ± 7.3	62/20	4-CV
2D FCN [Cai+17]	CT-82	80.3 ± 9.0	62/20	4-CV
2D FCN + RNN [Cai+17]	CT-82	82.3 ± 6.7	62/20	4-CV
Single Model 2D FCN [Zho+17]	CT-82	75.7 ± 10.5	62/20	4-CV
Multi-Model 2D FCN [Zho+17]	CT-82	82.2 ± 5.7	62/20	4-CV

training multiple, individual models.

Comparison to state-of-the-Art CT abdominal segmentation frameworks

The proposed architecture is evaluated on the public TCIA CT Pancreas benchmark to compare its performance with state-of-the-art methods. Initially, the models trained on CT-150 dataset are directly applied to CT-82 dataset to observe the applicability of the two models on different datasets. The corresponding results (BFT) are given in Table 7.3. U-Net model outperforms traditional atlas techniques [Wol+13] although it was trained on a disjoint dataset. Moreover, the attention model performs consistently better in pancreas segmentation across different datasets. These models are later fine-tuned (AFT) on a subset of TCIA dataset (61 train, 21 test). The output nodes corresponding to spleen and kidney are excluded from the output softmax computation, and the gradient updates are computed only for the background and pancreas labels. The results in Table 7.3 and 7.4 show improved performance compared to concatenated multi-model CNN approaches [Cai+17; Rot+18; Zho+17] due to additional training data and richer semantic information (e.g. spleen labels). Additionally, we trained the two models from scratch (SCR) with 61 training images randomly selected from the CT-82 dataset. Similar to the results on CT-150 dataset, AGs improve the segmentation accuracy and lower the surface distances ($p = .03$) due to increased recall rate of pancreas pixels ($p = .09$).

Table 7.5: Test results for standard scan plane detection. Number of initial filters is denoted by the postfix "-n". Time taken for forward (Fwd) and backward (Bwd) passes were recorded in milliseconds.

Method	Accuracy	F1	Precision	Recall	Fwd/Bwd (ms)	#Param
Sononet-8	0.969	0.899	0.878	0.922	1.36/2.60	0.16M
AG-Sononet-8	**0.977**	**0.922**	**0.916**	**0.929**	1.92/3.47	0.18M
Sononet-16	0.977	0.923	0.916	0.931	1.45/3.92	0.65M
AG-Sononet-16	**0.978**	**0.929**	**0.924**	**0.934**	1.94/5.13	0.70M
Sononet-32	0.979	0.931	0.924	**0.938**	2.40/6.72	2.58M
AG-Sononet-32	**0.980**	**0.933**	**0.931**	0.935	2.92/8.68	2.79M

Results from state-of-the-art CT pancreas segmentation models are summarised in Table 7.4 for comparison purposes. Since the models are trained on the same training dataset, this comparison gives an insight on how the attention model compares to the relevant literature. It is important to note that, post-processing (e.g. using conditional random field) is not utilised in our framework as the experiments mainly focus on quantification of performance improvement brought by AGs in an isolated setting. Similarly, residual and dense connections can be used as in [Gib+17] in conjunction with AGs to improve the segmentation results. In that regard, our 3D Attention U-Net model performs similar to the state-of-the-art, despite the input images are downsampled to lower resolution. More importantly, our approach significantly improves the results compared to single-model based segmentation frameworks (see Table 7.4). We do not require multiple CNN models to localise and segment object boundaries. Lastly, we performed 5-fold cross-validation on the CT-82 dataset using the Attention U-Net for a better comparison, which achieved 81.48 ± 6.23 DSC for pancreas labels.

7.3.4 2D fetal ultrasound image classification results

The dataset was split to training $(122,233)$, validation $(30,553)$ and testing $(38,243)$ frames on subject basis. For evaluation, we used macro-averaged precision, recall, F1, overall accuracy, the number of parameters and execution speed, summarised in Table 7.5.

In general, AG-Sononet improves the results over Sononet at all capacity levels. In particular,

Table 7.6: Class-wise performance for AG-Sononet-8. In bracket shows the improvement over Sononet-8. Bold highlights the improvement more than 0.02.

	Precision	Recall	F1
Brain (Cb.)	0.988 (-0.002)	0.982 (-0.002)	0.985 (-0.002)
Brain (Tv.)	0.980 (0.003)	0.990 (0.002)	0.985 (0.003)
Profile	0.953 **(0.055)**	0.962 (0.009)	0.958 **(0.033)**
Lips	0.976 **(0.029)**	0.956 (-0.003)	0.966 (0.013)
Abdominal	0.963 (0.011)	0.961 (0.007)	0.962 (0.009)
Kidneys	0.863 **(0.054)**	0.902 (0.003)	0.882 **(0.030)**
Femur	0.975 (0.019)	0.976 (-0.005)	0.975 (0.007)
Spine (Cor.)	0.935 **(0.049)**	0.979 (0.000)	0.957 **(0.026)**
Spine (Sag.)	0.936 **(0.055)**	0.979 (-0.012)	0.957 **(0.024)**
4CH	0.943 **(0.035)**	0.970 (0.007)	0.956 **(0.022)**
3VV	0.694 **(0.050)**	0.722 (-0.014)	0.708 **(0.021)**
RVOT	0.691 **(0.029)**	0.705 **(0.044)**	0.698 **(0.036)**
LVOT	0.925 **(0.022)**	0.933 **(0.027)**	0.929 **(0.024)**
Background	0.995 (-0.001)	0.992 (0.007)	0.993 (0.003)

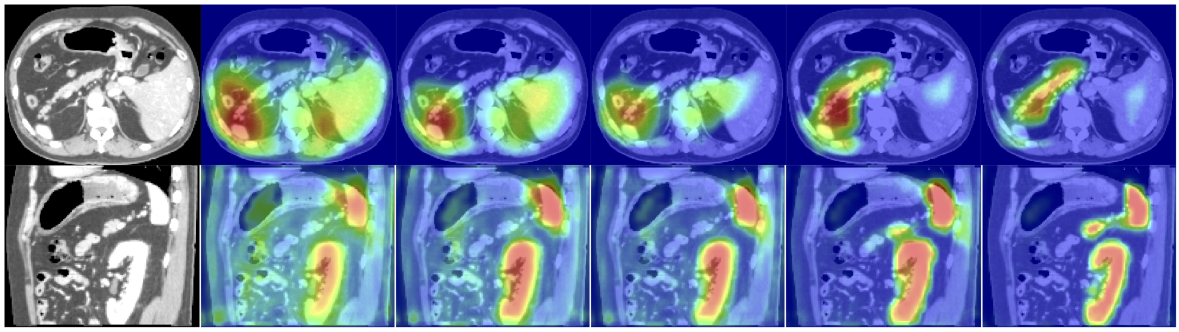

Figure 7.6: The figure shows the attention coefficients ($\alpha^{l_{s2}}$, $\alpha^{l_{s3}}$) across different training epochs (3, 6, 10, 60, 150). The images are extracted from sagittal and axial planes of a 3D abdominal CT scan from the testing dataset. The model gradually learns to focus on the pancreas, kidney, and spleen.

AG-Sononet achieves higher precision. AG-Sononet reduces false positive examples because the gating mechanism suppresses background noise and forces the network to make the prediction based on class-specific features. As the capacity of Sononet is increased, the gap between the methods is tightened, but we note that the performance of AG-Sononet is also close to the one of Sononet with double the capacity. In Table 7.6, we show the class-wise F1, precision and recall values for AG-Sononet-8, where the improvement over Sononet is indicated in brackets. We see that the precision increased by around 5% for kidney, profile and spines. For the most challenging cardiac views, we see on average 3% improvement for 4CH and 3VV ($p < 0.05$).

7.3.5 Attention map analysis

The attention coefficients of the proposed U-Net model, which are obtained from 3D-CT test images, are visualised with respect to training epochs (see Figure 7.6). We commonly observe that AGs initially have a uniform distribution and pass features at all spatial locations. This is gradually updated and localised towards the targeted organ boundaries. Additionally, at coarser scales AGs provide a rough outline of organs which are gradually refined at finer resolutions. Moreover, by training multiple AGs at each image scale, we observe that each AG learns to focus on a particular subset of organs.

Object localisation using attention Maps

With the proposed architecture, the localisation maps can obtained for almost no additional computational cost. In Figure 7.7, we show the attention maps of AG-Sononet across different subjects, together the red bounding box annotation generated using the attention maps. We see that the network consistently focuses on the object of interest, consistent with the blue ground truth annotation. We note, however, attention map outlines the discriminant region; in particular, it does not necessarily coincide with the entire object. Nevertheless, as it does not use guided backpropagation for localisation (a strategy in [Bau+16]), attention models are advantageous for real-time applications.

Finally, in Figure 7.7, we show the attention maps of AG-Sononet-FT across different subjects, together with the bounding box annotation generated using the attention maps. We see that the network consistently focuses on the object of interest, which indicates that the network indeed learnt the most important feature for each class. We note, however, attention map outlines the discriminant region; in particular, it does not necessarily coincide with the entire object. This behaviour makes sense because some part of object will appear in background label (i.e. when the ideal plane is not reached). Qualitatively, however, the bounding boxes well agree with the annotated ground truth. Most crucially, the attention map is obtained for almost no additional computational cost; In comparison, [Bau+16] requires guided backpropagation for localisation,

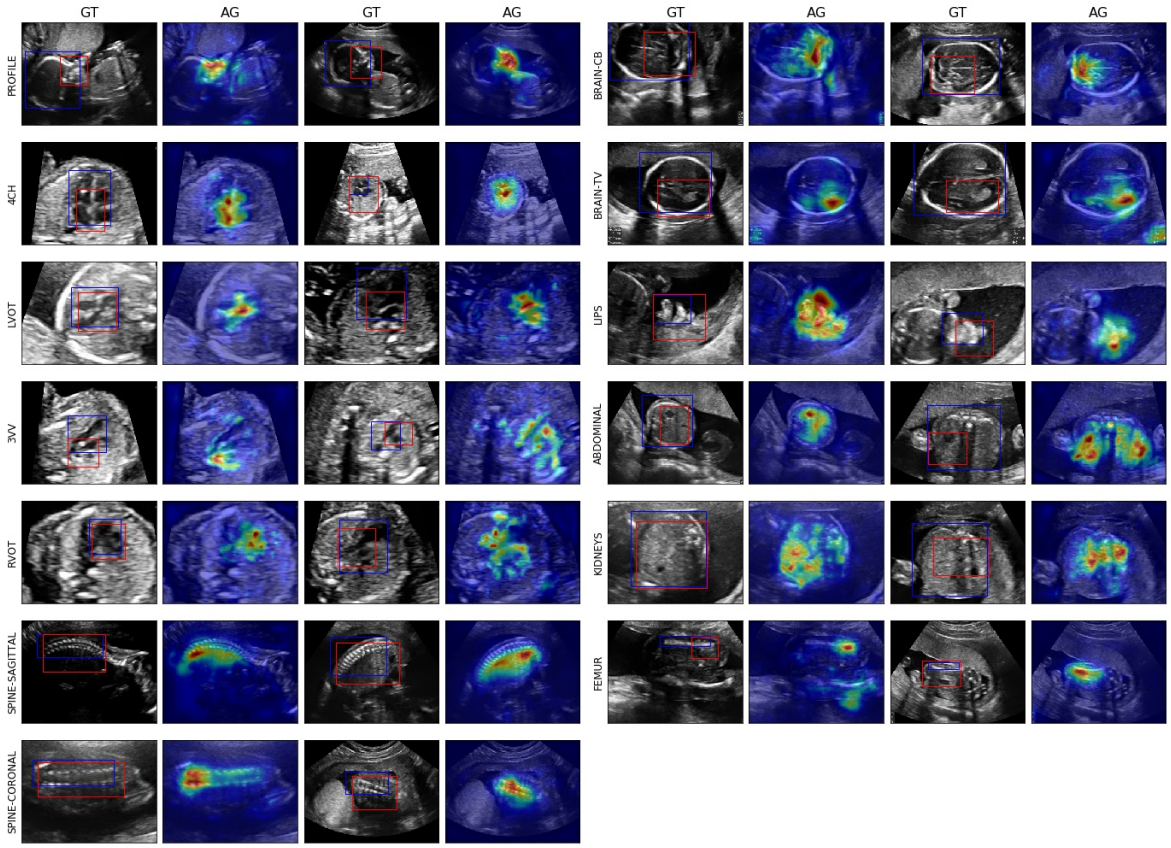

Figure 7.7: Examples of the obtained attention map and generated bounding boxes (red) from AG-Sononet-FT across different subjects. The ground truth annotation is shown in blue. The detected region highly agrees with the object of interest.

which limits the localisation speed. This highlights the advantage of attention model for the real-time applications.

7.4 Weakly supervised object localisation (WSL)

In [Bau+16], WSL was performed by exploiting the pixel-level saliency map obtained by guided-backpropagation, followed by ad-hoc procedure to extract bounding boxes. The same heuristics can be applied for the given network, however, owing to the attention map, we can devise a more efficient way of performing object localisation. In particular, we generate object location by simply: (1) blur the attention maps, (2) threshold the low activations, (3) perform connected-component analysis, (4) select a component that overlaps at each scale and (5) apply bounding

box around the selected components. In this heuristic, backpropagation is not required so it can be executed efficiently. We note, however, attention map outlines salient region used by the network to perform classification; in particular, it does not necessarily agree with the object of interest. This behaviour makes sense because some part of object will appear both in the class as well as background frame until the ideal plane is reached. Therefore, the quantitative result is shown in 7.7, however, the result is biased. We however define new metric called *Relative Correctness*, which is defined as 50% of maximum achievable IOU (due to bias). We see that in this metric, the method achieves very high results, indicating that it can detect relevant features of the object of interest in its proximity.

Table 7.7: WSL performance for the proposed strategy with AG-Sononet-16. Correctness (Cor.) is defined as $IOU > 0.5$. Relative Correctness (Rel.) is defined as $IOU > 0.5 \times \max(IOU_{class})$.

	IOU Mean (Std)	Cor. (%)	Rel. (%)
Brain (Cb.)	0.69 (0.11)	0.96	0.96
Brain (Tv.)	0.68 (0.12)	0.96	0.96
Profile	0.31 (0.08)	0.00	0.80
Lips	0.42 (0.18)	0.36	0.60
Abdominal	0.71 (0.10)	0.96	0.96
Kidneys	0.73 (0.13)	0.92	0.98
Femur	0.31 (0.11)	0.02	0.58
Spine (Cor.)	0.53 (0.13)	0.56	0.76
Spine (Sag.)	0.53 (0.11)	0.54	0.94
4CH	0.61 (0.14)	0.76	0.86
3VV	0.42 (0.14)	0.34	0.62
RVOT	0.56 (0.15)	0.70	0.76
LVOT	0.54 (0.15)	0.62	0.80

7.5 Discussion

In this work, we considered soft-attention mechanism and discussed how to incorporate this idea into segmentation and scan plane detection frameworks to better exploit local structures in CT abdominal and fetal ultrasound images. In particular, we highlighted several aspects: gridded attention mechanisms, a normalisation strategy for the attention map, and aggregation strategies. We empirically observed and reported that using soft-max as the activation function tends to generate a map that is sparsely activated and is overly sensitive to local

intensity changes. The latter is problematic as in ultrasound imaging, image quality is often low. In the classification setting, We found that dividing the activations by the sum of the activations helped generate attention map with larger contextual support. As demonstrated in the segmentation framework, Sigmoid function is a good alternative as it only normalises the range and allows more information to flow. However, we found that training is non-trivial due to the gradient saturation problem.

We noted that training the attention-mechanism was slightly more complex than the standard network architecture. In particular, we observed that the strategy employed to aggregate the attention maps at different scales affects both the learning of the attention mechanism itself and hence the performance. Having a loss term defined at each scale ensures that the network learns to attend at each scale. We observed that first training the network at each scale separately, followed by fine-tuning was the most stable approach to get the optimal performance.

There is a vast body of literature in machine learning exploring different gating architectures. For example, highway networks [GSS16] make use of residual connections around the gate block to allow better gradient back-propagation and slightly softer attention mechanisms. Although our segmentation experiments with residual connections have not provided any significant performance improvement, future work will focus on this aspect to obtain a better training behaviour.

Lastly, we note that the presented quantitative comparisons between the Attention 3D-Unet and state-of-the-art 2D cascaded models might not be sufficient enough to draw a final conclusion, as the proposed approach takes advantage of rich contextual information in all spatial dimensions. On the other hand, the 2D models utilise the high resolution information present in axial CT planes without any downsampling. We think that with the advent of improved GPU computation power and memory, larger capacity 3D-CT segmentation models can be trained with larger image grids without the need for image downsampling. In this regard, future research will focus more and more on deploying 3D models, and the performance of Attention U-Net can be further enhanced by utilising fine resolution input batches without any additional heuristics.

7.6 Conclusion

In this work we proposed a novel and modular attention gate model that can be easily incorporated into existing segmentation and classification architectures. Our approach can eliminate the necessity of applying an external object localisation model by implicitly learning to highlight salient regions in input images. Moreover, in a classification setting, AGs leverage the salient information to perform task adaptive feature pooling operation.

We applied the proposed attention model to standard scan plane detection during fetal ultrasound screening and showed that it improves overall results, especially precision, with much less parameters. This was done by generating the gating signal to pinpoint local as well as global information that is useful for the classification. Similarly, experimental results on CT segmentation task demonstrate that the proposed AGs are highly beneficial for tissue/organ identification and localisation. This is particularly true for variable small size organs such as the pancreas, and similar behaviour is observed in image classification tasks.

Additionally, AGs allow one to generate fine-grained attention map that can be exploited for object localisation. We envisage that the proposed soft-attention module could support explainable deep learning, which is a vital research area for medical imaging analysis.

Chapter 8

Conclusion

8.1 Summary

Medical imaging is an indispensable component of modern medical practice, however, due to high operational and maintenance costs as well as the current sub-optimal data processing pipelines, the medical devices remain less accessible to large population groups in the world. The objective of the thesis was to explore the approaches that can help us transition into smarter imaging protocols. In particular, three main problems were identified in this thesis. First one was the limitation in the current acquisition protocol. The second was the limitation in the current data processing pipeline, where we argued that combining acquisition, reconstruction and post-processing can allow us to optimise directly for the end-goal. The third problem addressed was the interpretability of automated methods in order to improve their reliability.

In this final chapter, we highlight the achievements of this thesis. The remainder of the chapter outlines the promising future directions and introduce some of our preliminary work that has already been done, in the hopes of reaching the goal of smarter imaging.

8.1.1 Achievements

Accelerated dynamic MR data reconstruction

Medical imaging devices are inherently complex and obtaining good image quality is a challenging task. For MRI, long acquisition time is required in order to produce high quality images. In particular, for cardiac MRI, one needs to acquire image of moving anatomy, which is extremely time-consuming and prolonged acquisition time is a burden for patients but also susceptible to motion artefact. Therefore, improving the acquisition speed for dynamic MRI is extremely beneficial.

In Chapter 4 and Chapter 5, we presented two approaches for accelerating dynamic MR image reconstruction, where we showed that the proposed method consistently outperforms state-of-the-art compressed sensing methods and is capable of preserving anatomical structure more faithfully up to 11-fold undersampling. In addition, both approaches enabled the reconstruction on GPU in less than 10 seconds. This is a clinically viable solution, compared to other compressed sensing methods, which can be time-consuming due to their iterative nature. This achievement should be able to improve the efficiency of the data acquisition, making the device more available, or able to provide this imaging technology in more time-critical environments.

Application-driven MRI: direct MR segmentation from undersampled k-space

In current medical imaging pipelines, there are four major distinct stages: acquisition, reconstruction, analysis and diagnosis. Typically, each step is optimised individually with respect to the final image quality. However, in many diagnostic scenarios, perfect reconstructions are not necessary as long as the images allow clinical practitioners to extract clinically relevant parameters. From this point of view, we argued that it is more efficient to optimise the process with the final goal in mind. we called this paradigm *application-driven imaging*.

In Chapter 6, we present a novel deep learning framework for extracting clinical parameters di-

rectly from undersampled cardiac MR data. We proposed two deep architectures, an end-to-end synthesis network and a latent feature interpolation network, to predict cardiac segmentation maps from extremely undersampled dynamic MRI data. In particular, we were able to reconstruct from less than 10 k-space lines per image. This highlights that for certain applications, even dramatically undersampled data have sufficient information to extract the clinical output. We envisage that the future of image acquisition with options to image specifically for output, and machine learning algorithms can aid making decisions about whether further in-depth imaging is required or not in a real-time fashion.

Interpretation of automated methods via attention models

It is expected that automated/semi-automated approaches in medical imaging analysis will be prevalent in the near future. However, it is increasingly important that one can gain sufficient understandings of how the automated methods work, such that if they fail, one knows how to deal with them. This aspect is especially important when we consider accelerated imaging or application-driven imaging, where multiple stages are combined holistically. In addition, understanding the method will not only provide the operators with the confidence in the methods but will also provide the scientist with the opportunity to improve the methods themselves. By unravelling how machine learning techniques handle information, it may facilitate the discovery of the new understandings of the problems.

In Chapter 7, we proposed the use of AG models for medical image analysis that automatically learns to focus on target structures of varying shapes and sizes. In particular, we applied AG networks for two tasks: CT segmentation and ultrasound scan plane detection and localisation. In CT segmentation, we saw that attention model can enhance the performance of segmenting small organs by ignoring the background information. Similar result was observed for ultrasound scan plane detection, where despite ultrasound having noisy backgrounds and low interpretability even for a clinician, the AG models were able to find consistent landmarks where it can perform plane classification. This ability was even utilised for weakly supervised

localisation, where despite not overlapping exactly with the anatomy, it often pointed to a subset of it, and we saw high consistency scores.

8.2 Limitations and future work

In this thesis, we have presented three methodologies which addressed each problem respectively. Nevertheless, there is still substantial work that needs be done before these methods could be deployed in practice. The last part of the thesis will highlight some of the existing issues and introduce our preliminary work which attempts to address them.

Prospective evaluation

For MRI reconstruction and direct segmentation, despite working with real imaging data, we still simulated the undersampling from fully sampled data. Ideally, the evaluation should be done with true data distribution, i.e. from raw undersampled data. In order to do so, we first need to extend our model to be able to cope with the common acquisition protocols.

There are two things we need to address in particular: non-Cartesian data and parallel imaging. While most imaging in a clinical setting uses Cartesian acquisition, non-Cartesian acquisition is often considered to be more robust to motion, as well as providing an efficient traversal of k-space and sufficiently incoherent artefact, which makes them well-suited for denoising type of reconstruction approaches. Recently, we have proposed a deep learning approach to handle non-Cartesian data in [Sch+19d] and achieved a state-of-the-art result for brain imaging. A sample reconstruction from the proposed approach is shown in Fig. 8.1. Similarly, we have recently extended our methods that appeared in Chapter 4 to parallel imaging reconstruction [Sch+19a; DSR19] for static images. The comparison with the state-of-the-art methods is shown in Fig. 8.2. The next step is to extend them for a dynamic imaging case so that we can study the realistic potential for these accelerated imaging techniques.

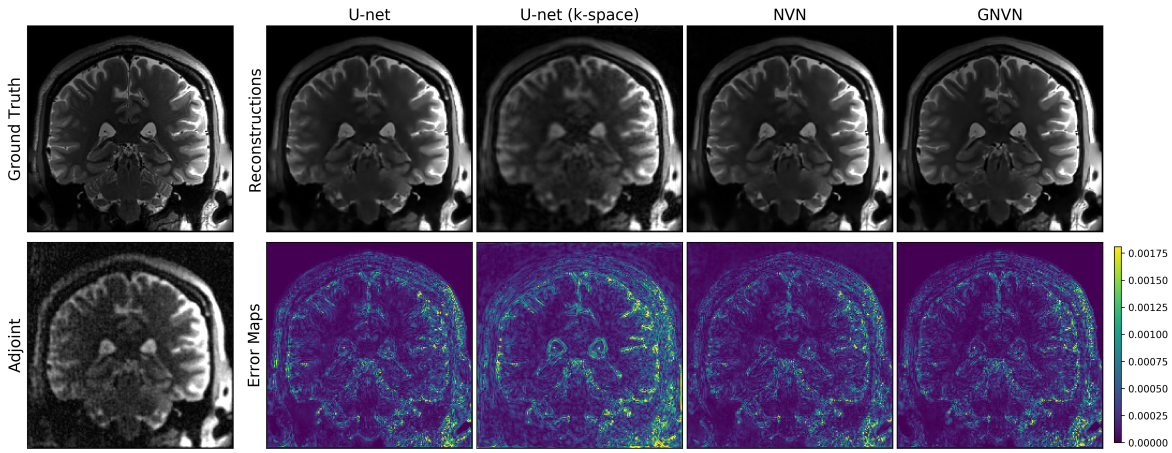

Figure 8.1: (Work from [Sch+19d]). The visualisation of the reconstructions of T2 weighted brain image, where images were undersampled with variable density sampling with AF 4.

Clinical evaluation

The second challenge is to prove the utility of the method with respect to clinical, diagnostic values. Currently, the evaluation is predominantly based on quantitative metrics, such as PSNR and SSIM for image reconstruction. For image segmentation, this is typically done by dice score. However, until clinical trials are performed, the true utility of these methods remain unclear. In Chapter 6, we evaluated the results with clinical values such as ejection fraction and ventricular volumes, which brings us slightly closer to studying the diagnostic relevance.

In [Sch+18d], we worked on cardiac diffusion tensor imaging reconstruction. Although only preliminary results were achieved, we performed the evaluation of the methods with respect to clinically relevant parameters such as fractional anisotropy, mean diffusivity and helix angle and show that the proposed DL reconstruction methods indeed provides the values more correlated with the ground truth measurements. The visualisation of these parameters is shown in Fig. 8.3.

Uncertainty estimation

As highlighted in Chapter 3, the current supervised learning framework focuses on obtaining the

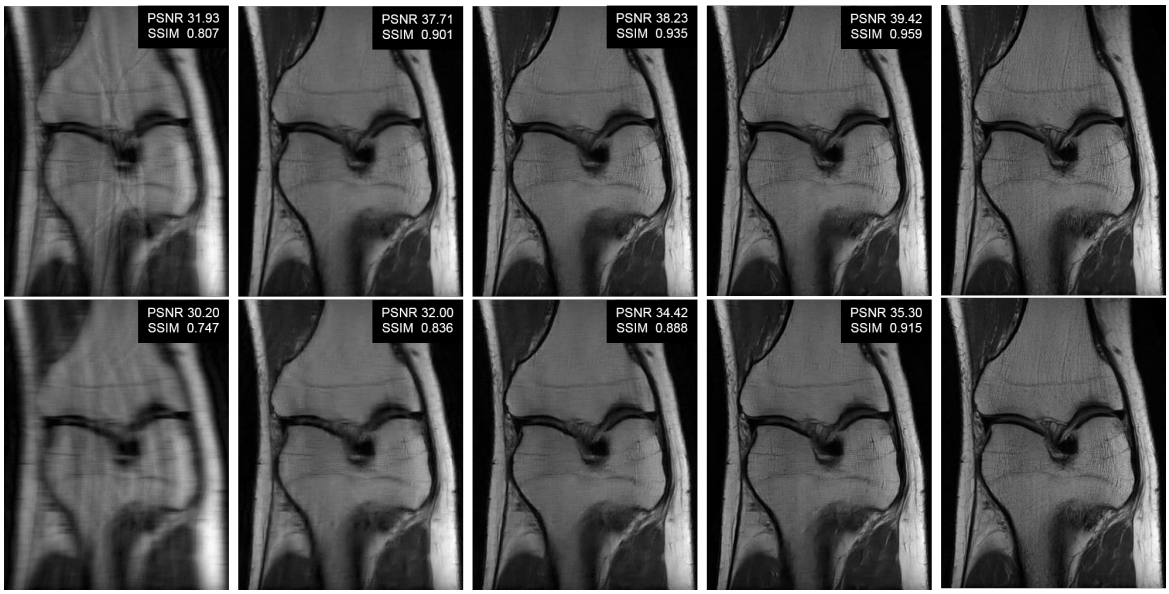

Figure 8.2: (Work from [DSR19]). The visual comparison of the parallel reconstruction of Cartesian undersampled knee-image for AF 4 (top) and 6 (bottom). From left to right: zero-filling results, Variational network [Ham+18], VS-Net (proposed), and ground truth.

best point estimate/prediction given input data. In practice, however, there are many sources of error in data acquisition or modelling. The best practice is to device a model that can accommodate with such sources of uncertainty. For example, for reconstruction task, if certain parts of the image are highly corrupted due to aliasing, the network should be less confident about its reconstruction. Rather than the model attempting to provide the best guess, it seems logical to equip these methods with the notion of *uncertainty estimate*. Uncertainty is somewhat complementary to attention-gate that we introduced in Chapter 7. While the latter provides the explainability of the how method works, the uncertainty estimates could potentially be used to explain when the method fails.

We have made a preliminary progress in applying *Bayesian deep learning* to MR image reconstruction. In [Sch+18d], we tried sub-network-dropout as a proxy for modelling the model uncertainty, which is also called *epistemic uncertainty*. In [Sch+18c], we modelled both epistemic and *aleatoric* uncertainty. The sample reconstructions and the uncertainty estimates are shown in Fig. 8.4. So far we observed that the uncertainty prediction is sensitive to edge, however, we also see correlation with the pixel intensity. Therefore, the output result does not seem to truly capture the underlying data/model uncertainty we expect to see. Recitfying

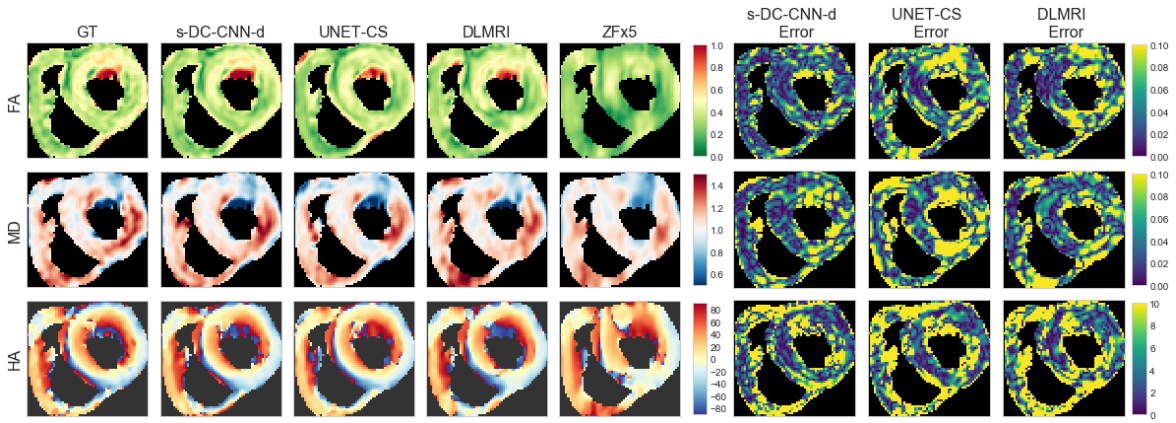

Figure 8.3: (Work from [Sch+18d]) The comparison of the diffusion tensor parameters and error maps from the proposed deep learning approach vs. baseline methods. From top to bottom: fractional anisotropy (FA), mean diffusivity (MD) ($10^{-3}\text{mm}^2\text{s}^{-1}$) and Helix-angle (HA) (degrees).

the uncertainty estimate is called *calibration*. We envisage the progress in this direction is important for the correct deployment of the Bayesian deep learning based models.

Theoretical guarantee

So far, deep learning has demonstrated its effectiveness for a wide range of applications. However, it is still an active area of research to investigate what makes them so work so well from a theoretical point of view. In other words, currently, most of the intuition underpinning the generalisation property of the deep neural networks is still empirical. Therefore, it is crucial to establish a theoretical framework that can explain its performance, so that one can characterise how these networks can fail in (realistic) the worst cases.

For the case of accelerated MR reconstruction, a variety of methods have been proposed, which generally fall in the category of k-space domain method, image domain methods, end-to-end learning method and iterative methods. For k-space methods, [YHC18] establishes a link between low-rank based approach (ALOHA) and deep networks. For the iterative methods, there are some preliminary work which attempts to quantify generalisation risk [Mar+19b]. AUTOMAP [Zhu+18a] is a prototypical approach which tries to learn the direct mapping from k-space to the output image space. Remarkably, despite having a large number of parameters,

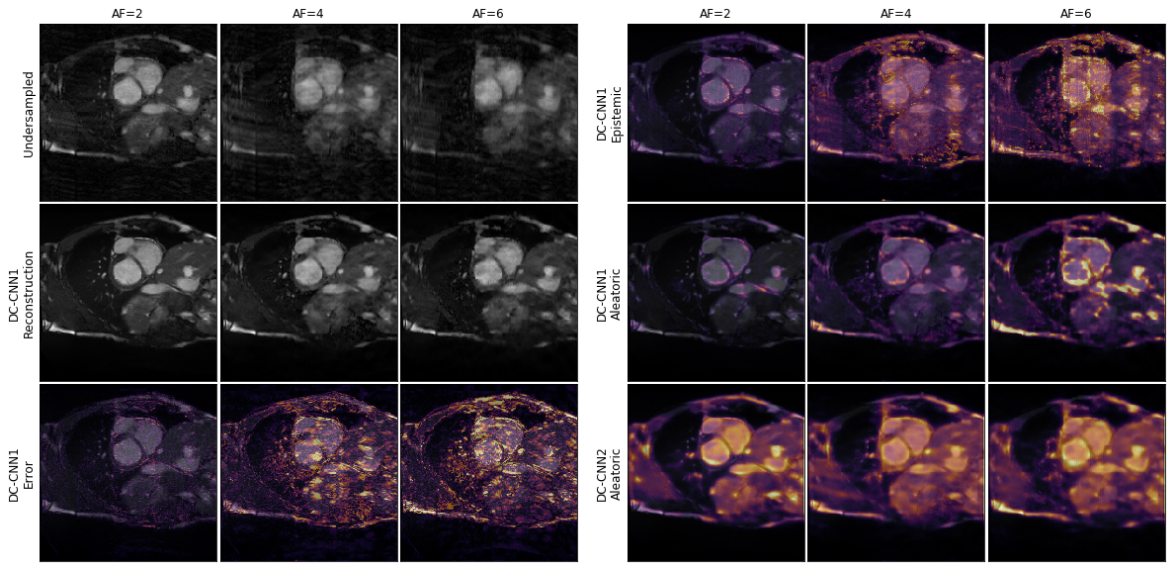

Figure 8.4: (Work from [Sch+18c]) The visualisation of epistemic and aleatoric uncertainty generated by two variants of the reconstruction networks, overlaid on the ground-truth image.

the method achieved robustness, however the theory behind its effectiveness is still non-existent. In [Sch+19b], we proposed a variant of approach called *dAUTOMAP*, which dramatically reduces the number of parameters by assuming the kernel-separability of the domain transform, and achieved a competitive performance to AUTOMAP (see Fig. 8.5 for the architecture). This highlights that there indeed is a lot of scope to improve the architectural design. However, it remains unknown how many parameters one needs to achieve the sufficient expressibility and generalisation given a task and the variation in the dataset. The problem of lack of theoretical development is not unique to MR reconstruction. For application-driven MRI, it is important that one can guarantee that the measured data actually contains sufficient information for one to obtain the clinical output. Therefore, a theoretical bound on how many k-space data sample is required would be an ideal addition to the proposed method. To this end, information theoretic approaches, such as data-processing inequality, might be one of the possible directions for theoretical analysis.

Domain adaptation, semi-supervised and unsupervised approaches

Currently in deep learning, the most successful branch is supervised learning. Supervised

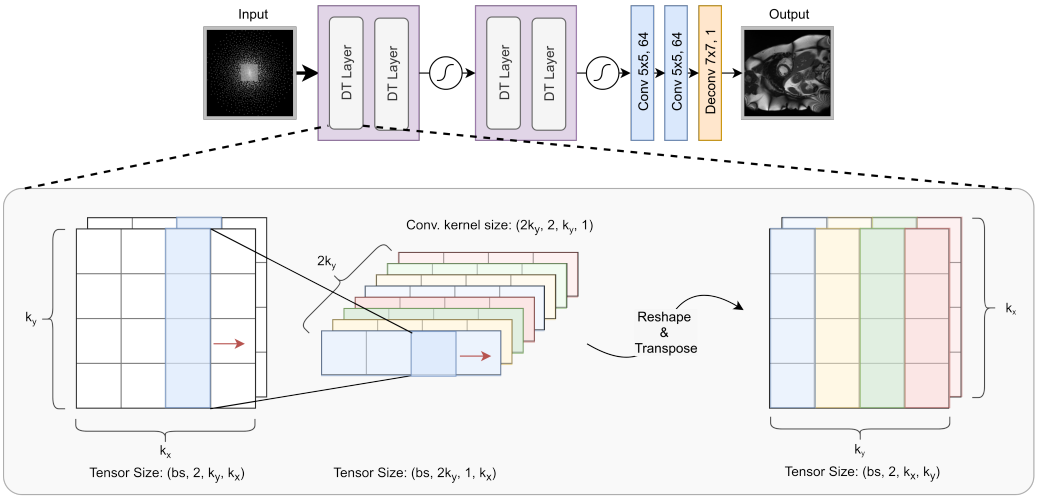

Figure 8.5: (Work from [Sch+19b]). A network architecture called *dAUTOMAP*. dAUTOMAP directly learns the domain transformation from raw k-space to an output image space. Unlike AUTOMAP [Zhu+18a], it exploits kernel-separability to significantly reduce the number of parameters.

learning was also utilised throughout this thesis. However, this is limiting in a sense, as it requires us to provide the model with a large amount of training data. This means that the application of supervised learning restricted to the cases where one can acquire true ground truth. This is problematic as in many cases, it is not possible to obtain perfect ground truth for inverse problems. In addition, whenever the domain between training and test environments deviates, one might need to retrain the algorithm with new examples, which is both time-consuming and inefficient. Therefore, in order to make supervised learning applicable to a wider range of problems, one must investigate how the learnt models can be generalised to an unseen distribution, solve the problem from limited number of examples, or solve the problem without requiring the ground truth data.

One promising approach is *domain adaptation and transfer learning*. These approaches bridge the gap between two domain distributions, called *domain shift*. For accelerated MR reconstruction, in [Ouy+19b] we proposed to train on a generic, large simulation-based dataset, then apply to other domain without any fine-tuning. We saw that the proposed approach had similar or better performance compared to training on real images. Studying the domain shift problem for different tasks (classification, segmentation and reconstruction) seems important

for us to better understand the subtle differences between these problems.

Another possible direction is unsupervised learning. In this formulation, the goal is to learn the prediction without having any ground truth data. Some work already exists, which employs Stein's unsupervised risk estimator (SURE). It was first used for image denoising, which was subsequently extended for image reconstruction. [ZSC18; Sha+18; Cha+19a; Leh+18]. Unsupservised learning approaches can be useful for the problems where it is difficult to simulate the forward model, or it is extremely difficult/time consuming to obtain correct ground truth data. It can be expected to see the deep learning community shift towards semi-supervised/unsupervised learning problems.

8.3 Final remark

The main goal of the work covered in thesis is to introduce deep learning approaches that can be used to improve medical imaging techniques holistically rather than focussing on medical image analysis and post-processing. While the techniques introduced in the thesis require more in-depth evaluation to be carried out in the future, we believe that the thesis serves as a great starting point for the broad spectrum of research in this direction. In our effort to accelerate the field, we make all the code from the main chapters publicly available. They can be found in the following Github pages:

- Ch 4 and 5: `https://github.com/js3611/Deep-MRI-Reconstruction`
- Ch. 6: `https://github.com/js3611/DirectCardiacSegmentation`
- Ch. 7: `https://github.com/ozan-oktay/Attention-Gatd-Networks`
- Ch. 8: `https://github.com/js3611/dAUTOMAP`

Publications

Journals

- **Schlemper, J.**, Caballero, J., Hajnal, J. V., Price, A. N., Rueckert, D. (2017). A deep cascade of convolutional neural networks for dynamic MR image reconstruction. IEEE transactions on Medical Imaging, 37(2), 491-503.
- Qin, C.†, **Schlemper, J.**†, Caballero, J., Price, A. N., Hajnal, J. V., Rueckert, D. (2018). Convolutional recurrent neural networks for dynamic MR image reconstruction. IEEE transactions on medical imaging, 38(1), 280-290.
- **Schlemper, J.**, Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis, 53, 197-207.

Conferences

- **Schlemper, J.**, Caballero, J., Hajnal, J. V., Price, A., Rueckert, D. (2017, June). A deep cascade of convolutional neural networks for MR image reconstruction. In International Conference on Information Processing in Medical Imaging (pp. 647-658). Springer, Cham.
- **Schlemper, J.**, Oktay, O., Bai, W., Castro, D.C., Duan, J., Qin, C., Hajnal, J.V. and Rueckert, D., 2018, September. Cardiac MR segmentation from undersampled k-space using deep latent representation learning. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 259-267). Springer, 2018.

- **Schlemper, J.**, Oktay, O., Chen, L., Matthew, J., Knight, C., Kainz, B., Glocker, B., Rueckert, D., Attention-gated networks for improving ultrasound scan plane detection., International Conference on Medical Imaging with Deep Learning, 2018.

- Oktay, O., **Schlemper, J.**, Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., Rueckert, D., Attention U-Net: learning where to look for the pancreas. International Conference on Medical Imaging with Deep Learning, 2018.

- **Schlemper, J.**†, Yang, G†, Ferreira, P., Scott, A., McGill, L., Khalique, Z., Gorodezky, M, Roehl, M., Keegan, J., Pennell, D., Firmin, D., Rueckert, D., Stochastic Deep Compressive Sensing for the Reconstruction of Diffusion Tensor Cardiac MRI. In:International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2018, pp. 295303.

- **Schlemper, J.**, Castro, D. C., Bai, W., Qin, C., Oktay, O., Duan, J., Price, A. N., Hajnal, J. V., Rueckert, D. Bayesian Deep Learning for Accelerated MR Image Reconstruction. In:International Workshop on Machine Learning for Medical Image Reconstruction. Springer. 2018, pp. 6471

- **Schlemper, J.**, Salehi, S. S. M.,and Kund, P., Lazarus, C., Dyvorne, H., Rueckert, D., Sofka, M., Nonuniform Variational Network: Deep Learning for Accelerated Nonuniform MR Image Reconstruction. In:International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2019.

- Duan, J.†, **Schlemper, J.**†, Rueckert, D. VS-Net: Variable spitting network for accelerated parallel MRI reconstruction. In:International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer., 2019.

Abstracts

- **Schlemper, J.**, Caballero, J., Hajnal, J. V., Price, A. N., Rueckert, D. (2017). A deep cascade of convolutional neural networks for MR image reconstruction. Abstract 0643,

25th Annual Meeting and Exhibition International Society of Magnetic Resonance in Medicine, 2017.

- **Schlemper, J.**, Duan, J., Ouyang, C., Qin, C., Caballero, J., Hajnal, J. V. and Rueckert, D. Data Consistency Networks for (Calibration-less) Accelerated Parallel MR Image Reconstruction. In:27th Annual Meeting, International Society for Magnetic Resonance in Medicine. 2019

- **Schlemper, J.**, Oksuz, I., Clough, J., Duan, J., King, A., Schnabel, J., Hajnal, J. V.,Rueckert, D. dAUTOMAP: Decomposing AUTOMAP to Achieve Scalability and Enhance Performance. In:27th Annual Meeting, International Society for Magnetic Resonance in Medicine. 2019

Bibliography

[Ach+10] Alin Achim et al. "Compressive sensing for ultrasound RF echoes using a-stable distributions". In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE. 2010, pp. 4304–4307.

[Akc+18] Mehmet Akcakaya et al. "Scan-specific robust artificial-neural-networks for k-space interpolation (RAKI) reconstruction: Database-free deep learning for fast imaging". In: *Magnetic Resonance in Medicine* 81.1 (2018).

[Ami+11] Md Faijul Amin et al. "Wirtinger calculus based gradient descent and Levenberg-Marquardt learning algorithms in complex-valued neural networks". In: *International Conference on Neural Information Processing*. Springer. 2011, pp. 550–559.

[AMJ17] Hemant Kumar Aggarwal, Merry P Mani, and Mathews Jacob. "MoDL: Model Based Deep Learning Architecture for Inverse Problems". In: *arXiv preprint arXiv:1712.02862* (2017).

[And+17] Peter Anderson et al. "Bottom-up and top-down attention for image captioning and VQA". In: *arXiv preprint arXiv:1707.07998* (2017).

[Ant+16] Joseph Antony et al. "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks". In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, pp. 1195–1200.

[AO17] Jonas Adler and Ozan Oktem. "Solving ill-posed inverse problems using iterative deep neural networks". In: *Inverse Problems* 33.2 (2017), p. 124007.

[AO18] Jonas Adler and Ozan Oktem. "Learned primal-dual reconstruction". In: *IEEE Transactions on Medical Imaging* (2018).

[Ars+18] Salim Arslan et al. "Graph Saliency Maps through Spectral Convolutional Networks: Application to Sex Classification with Brain Connectivity". In: *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities*. Springer, 2018, pp. 3–13.

[AVT14] Santiago Aja-Fernandez, Gonzalo Vegas-Sanchez-Ferrero, and Antonio Tristan-Vega. "Noise estimation in parallel MRI: GRAPPA and SENSE". In: *Magnetic resonance imaging* 32.3 (2014), pp. 281–290.

[Bai+18] Wenjia Bai et al. "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks". In: *Journal of Cardiovascular Magnetic Resonance* 20.1 (2018), p. 65.

[Ban16] The World Bank. *Data for High income, Upper middle income, Lower middle income, Low income.* 2016. URL: `https://data.worldbank.org/?locations=XD-XT-XN-XM`. (accessed: 11.06.2019).

[Bau+16] Christian F Baumgartner et al. "Real-Time Detection and Localisation of Fetal Standard Scan Planes in 2D Freehand Ultrasound". In: *arXiv preprint arXiv:1612.05601* (2016).

[BC17] Lorraine Brett and Lott Charles. *MRI Heart (Cardiac MRI).* 2017. URL: `https://www.insideradiology.com.au/cardiac-mri/`. (accessed: 17.07.2019).

[BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[Bel+19] Ghalib A Bello et al. "Deep-learning cardiac motion analysis for human survival prediction". In: *Nature machine intelligence* 1.2 (2019), p. 95.

[Ber+18] Olivier Bernard et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" In: *IEEE transactions on medical imaging* 37.11 (2018), pp. 2514–2525.

[BG09] Jurgen Braun and Michael Griebel. "On a constructive proof of Kolmogorov's superposition theorem". In: *Constructive approximation* 30.3 (2009), p. 653.

[Bif+19] Carlo Biffi et al. "3D High-Resolution Cardiac Segmentation Reconstruction from 2D Views using Conditional Variational Autoencoders". In: *arXiv preprint arXiv:1902.11000* (2019).

[Bis06] Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

[Bjo+18] Nils Bjorck et al. "Understanding batch normalization". In: *Advances in Neural Information Processing Systems.* 2018, pp. 7694–7705.

[BMS05] Oliver Bieri, M Markl, and Klaus Scheffler. "Analysis and compensation of eddy currents in balanced SSFP". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 54.1 (2005), pp. 129–137.

[Boo+01] John M Boone et al. "Dedicated breast CT: radiation dose and image quality evaluation". In: *Radiology* 221.3 (2001), pp. 657–667.

[Bow+17] Christopher Bowles et al. "Brain lesion segmentation through image synthesis and outlier detection". In: *NeuroImage: Clinical* 16 (2017), pp. 643–658.

[Bow+18] Christopher Bowles et al. "GANsfer Learning: Combining labelled and unlabelled data for GAN based data augmentation". In: *arXiv preprint arXiv:1811.10669* (2018).

[BRE17] Dmitry Batenkov, Yaniv Romano, and Michael Elad. "On the Global-Local Dichotomy in Sparsity Modeling". In: *Compressed Sensing and its Applications*. Springer, 2017, pp. 1–53.

[Bri+17] Denny Britz et al. "Massive exploration of neural machine translation architectures". In: *arXiv preprint arXiv:1703.03906* (2017).

[BRR12] John D Biglands, Aleksandra Radjenovic, and John P Ridgway. "Cardiovascular magnetic resonance physics for clinicians: Part II". In: *Journal of cardiovascular magnetic resonance* 14.1 (2012), p. 66.

[BUF07] Kai Tobias Block, Martin Uecker, and Jens Frahm. "Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 57.6 (2007), pp. 1086–1098.

[Bus+19] Aurelien Bustin et al. "Five-minute whole-heart coronary MRA with sub-millimeter isotropic resolution, 100% respiratory scan efficiency, and 3D-PROST reconstruction". In: *Magnetic resonance in medicine* 81.1 (2019), pp. 102–115.

[Cab+14a] Jose Caballero et al. "Application-Driven MRI: Joint Reconstruction and Segmentation from Undersampled MRI data". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*. Springer. 2014, pp. 106–113.

[Cab+14b] Jose Caballero et al. "Dictionary learning and time sparsity for dynamic MR data reconstruction". In: *IEEE Transactions on Medical Imaging* 33.4 (2014), pp. 979–994.

[Cai+17] Jinzheng Cai et al. "Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function". In: *MICCAI*. 2017.

[Cal88] Peter W Callen. "Ultrasonography in obstetrics and gynecology". In: (1988).

[Cam13] Stuart Campbell. "A short history of sonography in obstetrics and gynaecology". In: *Facts, views & vision in ObGyn* 5.3 (2013), p. 213.

[Can+18] Yigit B Can et al. "Learning to segment medical images with scribble-supervision alone". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 236–244.

[Can08] Emmanuel J Candes. "The restricted isometry property and its implications for compressed sensing". In: *Comptes rendus mathematique* 346.9-10 (2008), pp. 589–592.

[CBR13] Patricia Carreno-Moran, Julian Breeze, and Michael R Rees. "The Top Ten Cases in Cardiac MRI and the Most Important Differential Diagnoses". In: *Medical Imaging in Clinical Practice*. IntechOpen, 2013.

[CBR17] Liang Chen, Paul Bentley, and Daniel Rueckert. "Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks". In: *NeuroImage: Clinical* 15 (2017), pp. 633–643.

[Cer+18a] Juan J Cerrolaza et al. "3D fetal skull reconstruction from 2DUS via deep conditional generative networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 383–391.

[Cer+18b] Juan J Cerrolaza et al. "Deep learning with ultrasound physics for fetal skull segmentation". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 564–567.

[CH12] Chen Chen and Junzhou Huang. "Compressive sensing MRI with wavelet tree sparsity". In: *Advances in neural information processing systems*. 2012, pp. 1115–1123.

[Cha+09] LW Chan et al. "Volumetric (3D) imaging reduces inter-and intraobserver variation of fetal biometry measurements". In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 33.4 (2009), pp. 447–452.

[Cha+19a] Eunju Cha et al. "Boosting CNN beyond Label in Inverse Problems". In: *arXiv preprint arXiv:1906.07330* (2019).

[Cha+19b] Krishna Chaitanya et al. "Semi-supervised and Task-Driven Data Augmentation". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 29–41.

[Che+15] Hao Chen et al. "Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 507–514.

[Che+18] Liang Chen et al. "Drinet for medical image segmentation". In: *IEEE transactions on medical imaging* 37.11 (2018), pp. 2453–2462.

[Che+19a] Chen Chen et al. "Improving the generalizability of convolutional neural network-based segmentation on CMR images". In: *arXiv preprint arXiv:1907.01268* (2019).

[Che+19b] Chen Chen et al. "Learning Shape Priors for Robust Cardiac MR Segmentation from Multi-view Images". In: *arXiv preprint arXiv:1907.09983* (2019).

[Che+19c] Chen Chen et al. "Unsupervised Multi-modal Style Transfer for Cardiac MR Segmentation". In: *arXiv preprint arXiv:1908.07344* (2019).

[Chu+14] Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[Cio+17] Francesco Ciompi et al. "Towards automatic pulmonary nodule management in lung cancer screening with deep learning". In: *Scientific reports* 7 (2017), p. 46479.

[CLH18] Joseph Paul Cohen, Margaux Luck, and Sina Honari. "Distribution Matching Losses Can Hallucinate Features in Medical Image Translation". In: *arXiv preprint arXiv:1805.08841* (2018).

[Clo+19] James R Clough et al. "Global and Local Interpretability for Cardiac MRI Classification". In: *arXiv preprint arXiv:1906.06188* (2019).

[Cru+19] Gastao Cruz et al. "Rigid motion-corrected magnetic resonance fingerprinting". In: *Magnetic resonance in medicine* 81.2 (2019), pp. 947–961.

[CSJ11] Aladin Carovac, Fahrudin Smajlovic, and Dzelaludin Junuzovic. "Application of ultrasound in medicine". In: *Acta Informatica Medica* 19.3 (2011), p. 168.

[CSL16] Juan J Cerrolaza, Ronald M Summers, and Marius George Linguraru. "Soft multi-organ shape models via generalized PCA: A general framework". In: *MICCAI*. Springer. 2016, pp. 219–228.

[Dav83] Mark E Davison. "The ill-conditioned nature of the limited angle tomography problem". In: *SIAM Journal on Applied Mathematics* 43.2 (1983), pp. 428–448.

[DB98] Alexander H Delaney and Yoram Bresler. "Globally convergent edge-preserving regularized reconstruction: an application to limited-angle tomography". In: *IEEE Transactions on Image Processing* 7.2 (1998), pp. 204–221.

[Del+10] Benedicte MA Delattre et al. "Spiral demystified". In: *Magnetic resonance imaging* 28.6 (2010), pp. 862–881.

[Des+12] Anagha Deshmane et al. "Parallel MR imaging". In: *Journal of Magnetic Resonance Imaging* 36.1 (2012), pp. 55–72.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization". In: *Journal of Machine Learning Research* 12.Jul (2011), pp. 2121–2159.

[Don+06] David L Donoho et al. "Compressed sensing". In: *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306.

[Don+14] Chao Dong et al. "Learning a deep convolutional network for image super-resolution". In: *European Conference on Computer Vision*. Springer. 2014, pp. 184–199.

[Dou+16] Qi Dou et al. "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection". In: *IEEE Transactions on Biomedical Engineering* 64.7 (2016), pp. 1558–1567.

[Dou+18] Qi Dou et al. “Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss”. In: *arXiv preprint arXiv:1804.10916* (2018).

[Dro+16] Michal Drozdzal et al. “The importance of skip connections in biomedical image segmentation”. In: *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 179–187.

[DSR19] Jinming Duan, Jo Schlemper, and Daniel Rueckert. “VS-Net: Variable spitting network for accelerated parallel MRI reconstruction”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, (to appear).

[Dua+19] Jinming Duan et al. “Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach”. In: *IEEE transactions on medical imaging* (2019).

[EK12] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge university press, 2012.

[Elm90] Jeffrey L Elman. “Finding structure in time”. In: *Cognitive science* 14.2 (1990), pp. 179–211.

[Est+17] Andre Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature* 542.7639 (2017), p. 115.

[Fen+16] Li Feng et al. “XD-GRASP: golden-angle radial MRI with reconstruction of extra motion-state dimensions using compressed sensing”. In: *Magnetic resonance in medicine* 75.2 (2016), pp. 775–788.

[Fer+13] Pedro F Ferreira et al. “Cardiovascular magnetic resonance artefacts”. In: *Journal of Cardiovascular Magnetic Resonance* 15.1 (2013), p. 41.

[Fes07] Jeffrey A Fessler. “On NUFFT-based gridding for non-Cartesian MRI”. In: *Journal of Magnetic Resonance* 188.2 (2007), pp. 191–195.

[Fin92] Mathias Fink. “Time reversal of ultrasonic fields. I. Basic principles”. In: *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* 39.5 (1992), pp. 555–566.

[For+01] Kirsten PN Forbes et al. “PROPELLER MRI: clinical testing of a novel technique for quantification and compensation of head motion”. In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 14.3 (2001), pp. 215–222.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[GG15] Pooja Gaur and William A Grissom. “Accelerated MRI Thermometry by Direct Estimation of Temperature from Undersampled k-space Data”. In: *Magnetic Resonance in Medicine* 73.5 (2015), pp. 1914–1925.

[GHT17] Xiaohong W Gao, Rui Hui, and Zengmin Tian. "Classification of CT brain images based on deep learning networks". In: *Computer methods and programs in biomedicine* 138 (2017), pp. 49–56.

[Gib+17] Eli Gibson et al. "Towards image-guided pancreas and biliary endoscopy: Automatic multi-organ segmentation on abdominal CT with dense dilated networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 728–736.

[Gir+16] Rohit Girdhar et al. "Learning a Predictable and Generative Vector Representation for Objects". In: *Computer Vision – ECCV 2016*. Springer. 2016, pp. 484–499.

[Gon+18] Enhao Gong et al. "Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI". In: *Journal of Magnetic Resonance Imaging* 48.2 (2018), pp. 330–340.

[Goo+14] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[GR08] Michael J Gallagher and Gilbert L Raff. "Use of multislice CT for the evaluation of emergency room patients with chest pain: The so-called "Triple rule-out"". In: *Catheterization and cardiovascular interventions* 71.1 (2008), pp. 92–99.

[Gre+15] Karol Gregor et al. "DRAW: A Recurrent Neural Network For Image Generation". In: *Proceedings of The 32nd International Conference on Machine Learning*. 2015, pp. 1462–1471.

[GSS16] Klaus Greff, Rupesh K Srivastava, and Jurgen Schmidhuber. "Highway and residual networks learn unrolled iterative estimation". In: *arXiv preprint arXiv:1612.07771* (2016).

[Gua+18] Qingji Guan et al. "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification". In: *arXiv preprint arXiv:1801.09927* (2018).

[Guo+17] Yi Guo et al. "Direct estimation of tracer-kinetic parameter maps from highly undersampled brain dynamic contrast enhanced MRI". In: *Magnetic Resonance in Medicine* 78.4 (2017), pp. 1566–1578.

[Haa+99] E Mark Haacke et al. *Magnetic resonance imaging: physical principles and sequence design*. Vol. 82. Wiley-Liss New York: 1999.

[Hal13] Justin P Haldar. "Low-rank modeling of local k-space neighborhoods (LORAKS) for constrained MRI". In: *IEEE transactions on medical imaging* 33.3 (2013), pp. 668–681.

[Ham+13] Keijo Hamalainen et al. "Sparse tomography". In: *SIAM Journal on Scientific Computing* 35.3 (2013), B644–B665.

[Ham+17a] Kerstin Hammernik et al. "A deep learning architecture for limited-angle computed tomography reconstruction". In: *Bildverarbeitung fur die Medizin 2017*. Springer, 2017, pp. 92–97.

[Ham+17b] Kerstin Hammernik et al. "L2 or not L2: impact of loss function design for deep learning MRI reconstruction". In: *ISMRM 25th Annual Meeting*. 2017, p. 0687.

[Ham+18] Kerstin Hammernik et al. "Learning a Variational Network for Reconstruction of Accelerated MRI Data". In: *Magnetic resonance in medicine* 79 (2018), pp. 3055–3071.

[Han+18] Yoseob Han et al. "Deep learning with domain adaptation for accelerated projection-reconstruction MR". In: *Magnetic resonance in medicine* 80.3 (2018), pp. 1189–1205.

[Has+18] Seyed Raein Hashemi et al. "Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection". In: *IEEE Access* 7 (2018), pp. 1721–1735.

[Hau+09] Jorg Hausleiter et al. "Estimated radiation dose associated with cardiac CT angiography". In: *Jama* 301.5 (2009), pp. 500–507.

[Hau+19] Andreas Hauptmann et al. "Multi-Scale Learned Iterative Reconstruction". In: *arXiv preprint arXiv:1908.00936* (2019).

[HBO18] Mattias P Heinrich, Max Blendowski, and Ozan Oktay. "TernaryNet: Faster Deep Model Inference without GPUs for Medical 3D Segmentation using Sparse and Binary Convolutions". In: *arXiv preprint arXiv:1801.09449* (2018).

[HCA14] Junzhou Huang, Chen Chen, and Leon Axel. "Fast multi-contrast MRI reconstruction". In: *Magnetic resonance imaging* 32.10 (2014), pp. 1344–1352.

[He+15] Kaiming He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[He+16] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[Hea19] Siemens Healthineers. *Compressed Sensing - Beyond speed*. 2019. URL: `https://www.siemens-healthineers.com/magnetic-resonance-imaging/clinical-specialities/compressed-sensing`. (accessed: 10.07.2019).

[Hei+01] Robin M Heidemann et al. "VD-AUTO-SMASH imaging". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 45.6 (2001), pp. 1066–1074.

[HGE16] Ehsan Hosseini-Asl, Georgy Gimel'farb, and Ayman El-Baz. "Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network". In: *arXiv preprint arXiv:1607.00556* (2016).

[Hir+15] Akira Hirabayashi et al. "Compressed sensing MRI using sparsity induced from adjacent slice similarity". In: *2015 International Conference on Sampling Theory and Applications (SampTA)*. IEEE. 2015, pp. 287–291.

[HO17] Mattias P Heinrich and Ozan Oktay. "BRIEFnet: Deep Pancreas Segmentation using Binary Sparse Convolutions". In: *MICCAI*. Springer. 2017, pp. 329–337.

[Hou73] Godfrey N Hounsfield. "Computerized transverse axial scanning (tomography): Part 1. Description of system". In: *The British journal of radiology* 46.552 (1973), pp. 1016–1022.

[HS97] Sepp Hochreiter and Jurgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[HSS12] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent". In: *Cited on* 14 (2012).

[HSS17] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *arXiv preprint arXiv:1709.01507* (2017).

[Hua+17] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[Hub+09] Stefan Huber-Wagner et al. "Effect of whole-body CT during trauma resuscitation on survival: a retrospective, multicentre study". In: *The Lancet* 373.9673 (2009), pp. 1455–1461.

[HW83] Kenneth M Hanson and George W Wecksung. "Bayesian approach to limited-angle reconstruction in computed tomography". In: *JOSA* 73.11 (1983), pp. 1501–1509.

[HWW15] Yan Huang, Wei Wang, and Liang Wang. "Bidirectional recurrent convolutional networks for multi-frame super-resolution". In: *Advances in Neural Information Processing Systems*. 2015, pp. 235–243.

[HY18] Yoseob Han and Jong Chul Ye. "k-Space Deep Learning for Accelerated MRI". In: *arXiv preprint arXiv:1805.03779* (2018).

[HYY17] Yo Seob Han, Jaejun Yoo, and Jong Chul Ye. "Deep Learning with Domain Adaptation for Accelerated Projection Reconstruction MR". In: *arXiv preprint arXiv:1703.01135* (2017).

[HZ16] Justin P Haldar and Jingwei Zhuo. "P-LORAKS: Low-rank modeling of local k-space neighborhoods with parallel imaging data". In: *Magnetic resonance in medicine* 75.4 (2016), pp. 1499–1514.

[Iof17] Sergey Ioffe. "Batch renormalization: Towards reducing minibatch dependence in batch-normalized models". In: *Advances in neural information processing systems*. 2017, pp. 1945–1953.

[IS15] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167* (2015).

[JAF16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *European conference on computer vision*. Springer. 2016, pp. 694–711.

[Jak+98] Peter M Jakob et al. "AUTO-SMASH: a self-calibrating technique for SMASH imaging". In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 7.1 (1998), pp. 42–54.

[Jet+18] Saumya Jetley et al. "Learn to Pay Attention". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=HyzbhfWRW.

[Jin+17] Kyong Hwan Jin et al. "Deep convolutional neural network for inverse problems in imaging". In: *IEEE Transactions on Image Processing* 26.9 (2017), pp. 4509–4522.

[JLY15] Kyong Hwan Jin, Dongwook Lee, and Jong Chul Ye. "A novel k-space annihilating filter method for unification between compressed sensing and parallel MRI". In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2015, pp. 327–330.

[JLY16] Kyong Hwan Jin, Dongwook Lee, and Jong Chul Ye. "A general framework for compressed sensing and parallel MRI using annihilating filter based low-rank Hankel matrix". In: *IEEE Transactions on Computational Imaging* 2.4 (2016), pp. 480–495.

[Jou+17] Neige MY Journy et al. "Projected cancer risks potentially related to past, current, and future practices in paediatric CT in the United Kingdom, 1990–2020". In: *British journal of cancer* 116.1 (2017), p. 109.

[JYK07] Hong Jung, Jong Chul Ye, and Eung Yeop Kim. "Improved k-t BLAST and k-t SENSE using FOCUSS". In: *Physics in medicine and biology* 52.11 (2007), p. 3201.

[Kam+17] Konstantinos Kamnitsas et al. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation". In: *Medical image analysis* 36 (2017), pp. 61–78.

[Kam+18] K. Kamnitsas et al. "Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation". In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham, 2018, pp. 450–462.

[Kan+96] Masahiro Kaneko et al. "Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography." In: *Radiology* 201.3 (1996), pp. 798–802.

[KB15] Diederik Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *The International Conference on Learning Representations (ICLR)*. 2015.

[KCB16] Edward Kim, Miguel Corte-Real, and Zubair Baloch. "A deep semantic mobile application for thyroid cytopathology". In: *Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations*. Vol. 9789. International Society for Optics and Photonics. 2016, 97890A.

[KH16] Jeremy Kawahara and Ghassan Hamarneh. "Multi-resolution-tract CNN with hybrid pretrained and skin-lesion trained layers". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2016, pp. 164–171.

[KHY19] Shujaat Khan, Jaeyoung Huh, and Jong Chul Ye. "Deep Learning-based Universal Beamformer for Ultrasound Imaging". In: *arXiv preprint arXiv:1904.02843* (2019).

[KKK18] Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. "Fully Convolutional Multi-scale Residual DenseNets for Cardiac Segmentation and Automated Cardiac Diagnosis using Ensemble of Classifiers". In: *arXiv preprint arXiv:1801.05173* (2018).

[Kno+11a] Florian Knoll et al. "Adapted random sampling patterns for accelerated MRI". In: *Magnetic resonance materials in physics, biology and medicine* 24.1 (2011), pp. 43–50.

[Kno+11b] Florian Knoll et al. "Second order total generalized variation (TGV) for MRI". In: *Magnetic resonance in medicine* 65.2 (2011), pp. 480–491.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[Kul+16] Kuldeep Kulkarni et al. "Reconnet: Non-iterative reconstruction of images from compressively sensed measurements". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 449–458.

[KWW16] Jason Kuen, Zhenhua Wang, and Gang Wang. "Recurrent Attentional Networks for Saliency Detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3668–3677.

[L+95] Yann LeCun, Yoshua Bengio, et al. "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.

[Laz+19] Carole Lazarus et al. "SPARKLING: variable-density k-space filling curves for accelerated T2*-weighted MRI". In: *Magnetic Resonance in Medicine* (2019).

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), p. 436.

[Led+17] Christian Ledig et al. "Photo-realistic single image super-resolution using a generative adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4681–4690.

[Lee+15] Chen-Yu Lee et al. "Deeply-supervised nets". In: *Artificial Intelligence and Statistics*. 2015, pp. 562–570.

[Leh+18] Jaakko Lehtinen et al. “Noise2noise: Learning image restoration without clean data”. In: *arXiv preprint arXiv:1803.04189* (2018).

[LH15] Ming Liang and Xiaolin Hu. “Recurrent convolutional neural network for object recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3367–3375.

[Lia+09] Dong Liang et al. “Accelerating SENSE using compressed sensing”. In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 62.6 (2009), pp. 1574–1584.

[Lia+17] Fangzhou Liao et al. “Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network”. In: *arXiv preprint arXiv:1711.08324* (2017).

[Lin+11] Sajan Goud Lingala et al. “Accelerated dynamic MRI exploiting sparsity and low-rank structure: K-t SLR”. In: *IEEE Transactions on Medical Imaging* 30.5 (2011), pp. 1042–1054.

[Lit+17] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: *Medical image analysis* 42 (2017), pp. 60–88.

[Liu+17] Jun Liu et al. “Global context-aware attention lstm networks for 3d action recognition”. In: *CVPR*. 2017.

[LKG19] Zeju Li, Konstantinos Kamnitsas, and Ben Glocker. “Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation”. In: *arXiv preprint arXiv:1907.10982* (2019).

[LL19] Alexander Selvikvag Lundervold and Arvid Lundervold. “An overview of deep learning in medical imaging focusing on MRI”. In: *Zeitschrift fur Medizinische Physik* 29.2 (2019), pp. 102–127.

[LL91] JF Larsen and T Larsen. “Unskilled ultrasound scanning in gynecology and obstetrics can lead to serious risks”. In: *Ugeskrift for laeger* 153.31 (1991), pp. 2187–2188.

[LN07] David J Larkman and Rita G Nunes. “Parallel magnetic resonance imaging”. In: *Physics in Medicine & Biology* 52.7 (2007), R15.

[LPF13] Herve Liebgott, Remy Prost, and Denis Friboulet. “Pre-beamformed RF signal reconstruction in medical ultrasound using compressive sensing”. In: *Ultrasonics* 53.2 (2013), pp. 525–533.

[LPM15] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. “Effective approaches to attention-based neural machine translation”. In: *arXiv preprint arXiv:1508.04025* (2015).

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[Lu+16] Jiasen Lu et al. "Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning". In: *CoRR* abs/1612.01887 (2016). arXiv: 1612.01887. URL: http://arxiv.org/abs/1612.01887.

[Lus+05] Michael Lustig et al. "Faster imaging with randomly perturbed, under-sampled spirals and l1 reconstruction". In: *Proceedings of the 13th annual meeting of ISMRM, Miami Beach*. Citeseer. 2005, p. 685.

[Lus+08] Michael Lustig et al. "Compressed sensing MRI". In: *IEEE signal processing magazine* 25.2 (2008), pp. 72–82.

[LYY17] Dongwook Lee, Jaejun Yoo, and Jong Chul Ye. "Deep artifact learning for compressed sensing and parallel MRI". In: *arXiv preprint arXiv:1703.01120* (2017).

[M+14] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. "Recurrent models of visual attention". In: *Advances in neural information processing systems*. 2014, pp. 2204–2212.

[Mad+18] Ali Madani et al. "Fast and accurate view classification of echocardiograms using deep learning". In: *npj Digital Medicine* 1.1 (2018), p. 6.

[Mag18] Maglab. *MRI: A Guided Tour*. Last modified on 20 September 2018. URL: https://nationalmaglab.org/education/magnet-academy/learn-the-basics/stories/mri-a-guided-tour. (accessed: 10.07.2019).

[Mai+19] Andreas Maier et al. "A gentle introduction to deep learning in medical image processing". In: *Zeitschrift fur Medizinische Physik* 29.2 (2019), pp. 86–101.

[Mar+17] Morteza Mardani et al. "Recurrent generative adversarial networks for proximal learning and automated compressive image recovery". In: *arXiv preprint arXiv:1711.10046* (2017).

[Mar+19a] Morteza Mardani et al. "Deep Generative Adversarial Neural Networks for Compressive Sensing MRI". In: *IEEE transactions on medical imaging* 38.1 (2019), pp. 167–179.

[Mar+19b] Morteza Mardani et al. "Degrees of Freedom Analysis of Unrolled Neural Networks". In: *arXiv preprint arXiv:1906.03742* (2019).

[Mat+13] John D Mathews et al. "Cancer risk in 680 000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians". In: *Bmj* 346 (2013), f2360.

[McC+16] Michael T McCann et al. "Fast 3D reconstruction method for differential phase contrast X-ray CT". In: *Optics express* 24.13 (2016), pp. 14564–14581.

[Men+19] Qingjie Meng et al. "Weakly Supervised Estimation of Shadow Confidence Maps in Fetal Ultrasound Imaging". In: *IEEE transactions on medical imaging* (2019).

[Met91] T Mets. "Clinical ultrasound in developing countries". In: *The Lancet* 337.8737 (1991), p. 358.

[MF15] Madison G McGaffin and Jeffrey A Fessler. "Alternating dual updates algorithm for X-ray CT reconstruction on the GPU". In: *IEEE transactions on computational imaging* 1.3 (2015), pp. 186–199.

[Mia+16] Xin Miao et al. "Accelerated cardiac cine MRI using locally low rank and finite difference constraints". In: *Magnetic resonance imaging* 34.6 (2016), pp. 707–714.

[MM03] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[MNA16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE. 2016, pp. 565–571.

[Mur+12] Mark Murphy et al. "Fast *ell*_1-SPIRiT Compressed Sensing Parallel Imaging MRI: Scalable Parallel Implementation and Clinically Feasible Runtime". In: *IEEE transactions on medical imaging* 31.6 (2012), pp. 1250–1262.

[Nai+90] David P Naidich et al. "Low-dose CT of the lungs: preliminary observations." In: *Radiology* 175.3 (1990), pp. 729–731.

[Nes83] Y. Nesterov. "A method of solving a convex programming problem with convergence rate $O(1/sqr(k))$". In: *Soviet Mathematics Doklady*. Vol. 27. 1983, pp. 372–376.

[NHK16] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. "Dual Attention Networks for Multimodal Reasoning and Matching". In: *CoRR* abs/1611.00471 (2016). arXiv: `1611.00471`. URL: `http://arxiv.org/abs/1611.00471`.

[NHS15] NHS Screening Programmes. *Fetal Anomaly Screen Programme Handbook*. London, UK: NHS, 2015.

[Nik+04] Konstantin Nikolaou et al. "Advances in cardiac CT imaging: 64-slice scanner". In: *The international journal of cardiovascular imaging* 20.6 (2004), pp. 535–540.

[Nis] Dwight George Nishimura. "Principles of magnetic resonance imaging". In: ().

[Nov+99] Robert A Novelline et al. "Helical CT in emergency radiology". In: *Radiology* 213.2 (1999), pp. 321–339.

[NQ14] Sherif F Nagueh and Miguel A Quinones. "Important advances in technology: echocardiography". In: *Methodist DeBakey cardiovascular journal* 10.3 (2014), p. 146.

[OCS15] Ricardo Otazo, Emmanuel Candes, and Daniel K. Sodickson. "Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components". In: *Magnetic Resonance in Medicine* 73.3 (2015), pp. 1125–1136.

[Oda+17] Masahiro Oda et al. "3D FCN Feature Driven Regression Forest-Based Pancreas Localization and Segmentation". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, 222–==9230.

[Oks+18] Ilkay Oksuz et al. "Deep learning using K-space based data augmentation for automated cardiac MR motion artefact detection". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 250–258.

[Oks+19] Ilkay Oksuz et al. "Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning". In: *Medical image analysis* 55 (2019), pp. 136–147.

[Okt+17] Ozan Oktay et al. "Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation". In: *IEEE transactions on medical imaging* 37.2 (2017), pp. 384–395.

[Org+98] World Health Organization et al. "Training in diagnostic ultrasound: essentials, principles and standards: report of a WHO study group". In: (1998).

[Org17] World Health Organization. *Global Atlas of Medical Devices. (Who Medical Device Technical Series)*. 2017. URL: `https://apps.who.int/medicinedocs/en/m/abstract/Js23215en/`. (accessed: 11.06.2019).

[Org18] World Health Organization. *The top 10 causes of death*. May 2018. URL: `https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death`. (accessed: 11.06.2019).

[Ouy+19a] Cheng Ouyang et al. "Data Efficient Unsupervised Domain Adaptation for Cross-Modality Image Segmentation". In: *arXiv preprint arXiv:1907.02766* (2019).

[Ouy+19b] Cheng Ouyang et al. "Generalising Deep Learning MRI Reconstruction across Different Domains". In: *arXiv preprint arXiv:1902.10815* (2019).

[Pas+17] Adam Paszke et al. "Automatic differentiation in pytorch". In: (2017).

[Pau17] John Pauly. *EE369C: Medical Image Reconstruction*. 2017. URL: `http://web.stanford.edu/class/ee369c/`. (accessed: 10.07.2019).

[Paw+19] Nick Pawlowski et al. "Needles in Haystacks: On Classifying Tiny Objects in Large Images". In: *arXiv preprint arXiv:1908.06037* (2019).

[Pay+17] Christian Payer et al. "Multi-label Whole Heart Segmentation Using CNNs and Anatomical Label Configurations". In: *STACOM*. Springer. 2017, pp. 190–198.

[Pea+12] Mark S Pearce et al. "Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study". In: *The Lancet* 380.9840 (2012), pp. 499–505.

[Pei+16] Wenjie Pei et al. "Temporal Attention-Gated Model for Robust Sequence Classification". In: *CoRR* abs/1612.00385 (2016). arXiv: 1612.00385. URL: http://arxiv.org/abs/1612.00385.

[Pes+17] Emanuele Pesce et al. "Learning to detect chest radiographs containing lung nodules using visual attention networks". In: *arXiv preprint arXiv:1712.00996* (2017).

[Pet+16] Steffen E. Petersen et al. "UK Biobank's Cardiovascular Magnetic Resonance Protocol". In: *Journal of Cardiovascular Magnetic Resonance* 18.1 (Feb. 2016), p. 8. ISSN: 1532-429X. DOI: 10.1186/s12968-016-0227-4.

[Pip99] James G Pipe. "Motion correction with PROPELLER MRI: application to head motion and free-breathing cardiac imaging". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 42.5 (1999), pp. 963–969.

[PJ15] Sunrita Poddar and Mathews Jacob. "Dynamic MRI using smoothness regularization on manifolds (SToRM)". In: *IEEE transactions on medical imaging* 35.4 (2015), pp. 1106–1115.

[PL14] Xi Peng and Dong Liang. "MR image reconstruction with convolutional characteristic constraint (CoCCo)". In: *IEEE Signal Processing Letters* 22.8 (2014), pp. 1184–1188.

[Pra+14] Somnath J Prabhu et al. "Ultrasound artifacts: classification, applied physics with illustrations, and imaging appearances". In: *Ultrasound quarterly* 30.2 (2014), pp. 145–157.

[Pru+01] Klaas P Pruessmann et al. "Advances in sensitivity encoding with arbitrary k-space trajectories". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 46.4 (2001), pp. 638–651.

[Pru+99] Klaas P Pruessmann et al. "SENSE: sensitivity encoding for fast MRI". In: *Magnetic resonance in medicine* 42.5 (1999), pp. 952–962.

[Qin+18a] Chen Qin et al. "Joint learning of motion estimation and segmentation for cardiac MR image sequences". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 472–480.

[Qin+18b] Chen Qin et al. "Joint motion estimation and segmentation from undersampled cardiac MR image". In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2018, pp. 55–63.

[Qin+19] Chen Qin et al. "Convolutional recurrent neural networks for dynamic MR image reconstruction". In: *IEEE transactions on medical imaging* 38.1 (2019), pp. 280–290.

[QJ16] Tran Minh Quan and Won-Ki Jeong. "Compressed sensing reconstruction of dynamic contrast enhanced MRI using GPU-accelerated convolutional sparse coding". In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2016, pp. 518–521.

[QNJ18] Tran Minh Quan, Thanh Nguyen-Duc, and Won-Ki Jeong. "Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss". In: *IEEE transactions on medical imaging* 37.6 (2018), pp. 1488–1497.

[Qui+10] Celine Quinsac et al. "Compressed sensing of ultrasound images: Sampling of spatial and frequency domains". In: *2010 IEEE Workshop On Signal Processing Systems*. IEEE. 2010, pp. 231–236.

[R+88] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. "Learning representations by back-propagating errors". In: *Cognitive modeling* 5.3 (1988), p. 1.

[Raj+16] Martin Rajchl et al. "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks". In: *IEEE transactions on medical imaging* 36.2 (2016), pp. 674–683.

[Ras+95] Volker Rasche et al. "Continuous radial data acquisition for dynamic MRI". In: *Magnetic resonance in medicine* 34.5 (1995), pp. 754–761.

[RB11] Saiprasad Ravishankar and Yoram Bresler. "MR image reconstruction from highly undersampled k-space data by dictionary learning". In: *IEEE transactions on medical imaging* 30.5 (2011), pp. 1028–1041.

[Ren+15] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems*. 2015, pp. 91–99.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[Ric+17] JH Rick Chang et al. "One Network to Solve Them All–Solving Linear Inverse Problems Using Deep Projection Models". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5888–5897.

[Rid10] John P Ridgway. "Cardiovascular magnetic resonance physics for clinicians: part I". In: *Journal of cardiovascular magnetic resonance* 12.1 (2010), p. 71.

[Rie+88] Stephen J Riederer et al. "MR fluoroscopy: technical feasibility". In: *Magnetic resonance in medicine* 8.1 (1988), pp. 1–15.

[RNZ18] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. "Deep learning for medical image processing: Overview, challenges and the future". In: *Classification in BioApps*. Springer, 2018, pp. 323–350.

[Roe+90] Peter B Roemer et al. "The NMR phased array". In: *Magnetic resonance in medicine* 16.2 (1990), pp. 192–225.

[Ros61] Frank Rosenblatt. *Principles of neurodynamics. perceptrons and the theory of brain mechanisms*. Tech. rep. Cornell Aeronautical Lab Inc Buffalo NY, 1961.

[Rot+16] Holger Roth et al. *Data From Pancreas-CT. The Cancer Imaging Archive*. 2016. URL: http://doi.org/10.7937/K9/TCIA.2016.tNB1kqBU (visited on 03/22/2018).

[Rot+17] Holger R Roth et al. "Hierarchical 3D fully convolutional networks for multi-organ segmentation". In: *arXiv preprint arXiv:1704.06382* (2017).

[Rot+18] Holger R. Roth et al. "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation". In: *Medical Image Analysis* 45 (2018), pp. 94–107. ISSN: 1361-8415.

[Rud16] Sebastian Ruder. "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747* (2016).

[RZ16] Mengye Ren and Richard S. Zemel. "End-to-End Instance Segmentation and Counting with Recurrent Attention". In: *CoRR* abs/1605.09410 (2016). arXiv: 1605.09410. URL: http://arxiv.org/abs/1605.09410.

[S+03] Patrice Y Simard, David Steinkraus, John C Platt, et al. "Best practices for convolutional neural networks applied to visual document analysis." In: *Icdar*. Vol. 3. 2003. 2003.

[S+16] Jian Sun, Huibin Li, Zongben Xu, et al. "Deep ADMM-Net for compressive sensing MRI". In: *Advances in neural information processing systems*. 2016, pp. 10–18.

[San+18] Shibani Santurkar et al. "How does batch normalization help optimization?" In: *Advances in Neural Information Processing Systems*. 2018, pp. 2483–2493.

[Sar+17] Saman Sarraf et al. "DeepAD: Alzheimer's Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI". In: *bioRxiv* (2017). DOI: 10.1101/070441. URL: https://www.biorxiv.org/content/early/2017/01/14/070441.

[SB95] PC Seynaeve and JI Broos. "The history of tomography". In: *Journal belge de radiologie* 78.5 (1995), pp. 284–288.

[Sch+17] Jo Schlemper et al. "A Deep Cascade of Convolutional Neural Networks for MR Image Reconstruction". In: *International Conference on Information Processing in Medical Imaging*. 2017, pp. 647–658.

[Sch+18a] Jo Schlemper et al. "A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction". In: *IEEE Transactions on Medical Imaging* 37.2 (2018).

[Sch+18b] Jo Schlemper et al. "Attention-gated networks for improving ultrasound scan plane detection". In: *arXiv preprint arXiv:1804.05338* (2018).

[Sch+18c] Jo Schlemper et al. "Bayesian Deep Learning for Accelerated MR Image Reconstruction". In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2018, pp. 64–71.

[Sch+18d] Jo Schlemper et al. "Stochastic Deep Compressive Sensing for the Reconstruction of Diffusion Tensor Cardiac MRI". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 295–303.

[Sch+19a] Jo Schlemper et al. "Data Consistency Networks for (Calibration-less) Accelerated Parallel MR Image Reconstruction". In: *27th Annual Meeting, International Society for Magnetic Resonance in Medicine*. 2019.

[Sch+19b] Jo Schlemper et al. "dAUTOMAP: Decomposing AUTOMAP to Achieve Scalability and Enhance Performance". In: *27th Annual Meeting, International Society for Magnetic Resonance in Medicine*. 2019.

[Sch+19c] Jo Schlemper et al. "Deep Hashing using Entropy Regularised Product Quantisation Network". In: *arXiv preprint arXiv:1902.03876* (2019).

[Sch+19d] Jo Schlemper et al. "Nonuniform Variational Network: Deep Learning for Accelerated Nonuniform MR Image Reconstruction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, (to appear).

[Sch16] Harvard Medical School. *Do CT scans cause cancer?* 2016. URL: `https://www.health.harvard.edu/staying-healthy/do-ct-scans-cause-cancer`. (accessed: 18.07.2019).

[SEG17a] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. "Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging". In: *IEEE transactions on medical imaging* 36.11 (2017), pp. 2319–2330.

[SEG17b] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. "Tversky loss function for image segmentation using 3D fully convolutional deep networks". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2017, pp. 379–387.

[Sei+18] Maximilian Seitzer et al. "Adversarial and perceptual refinement for compressed sensing MRI reconstruction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 232–240.

[Sha+18] Fahad Shamshad et al. "Leveraging Deep Stein's Unbiased Risk Estimator for Unsupervised X-ray Denoising". In: *arXiv preprint arXiv:1811.12488* (2018).

[She+15] Wei Shen et al. “Multi-scale convolutional neural networks for lung nodule classification”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2015, pp. 588–599.

[She+17] Tao Shen et al. “Disan: Directional self-attention network for rnn/cnn-free language understanding”. In: *arXiv preprint arXiv:1709.04696* (2017).

[Shi+14] Peter J Shin et al. “Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion”. In: *Magnetic resonance in medicine* 72.4 (2014), pp. 959–970.

[Shi+16] Wenzhe Shi et al. “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1874–1883.

[Sin+18] Matthew Sinclair et al. “Human-level performance on automatic head biometrics in fetal ultrasound using fully convolutional neural networks”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 714–717.

[SM97] Daniel K Sodickson and Warren J Manning. “Simultaneous acquisition of spatial harmonics (SMASH): fast imaging with radiofrequency coil arrays”. In: *Magnetic resonance in medicine* 38.4 (1997), pp. 591–603.

[SNS16] Atsushi Saito, Shigeru Nawano, and Akinobu Shimizu. “Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs”. In: *Medical image analysis* 28 (2016), pp. 46–65.

[SP08] Emil Y Sidky and Xiaochuan Pan. “Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization”. In: *Physics in Medicine & Biology* 53.17 (2008), p. 4777.

[SP15] M Soleimani and T Pengpen. *Introduction: a brief overview of iterative algorithms in x-ray computed tomography*. 2015.

[Spa+16] Fabio Alexandre Spanhol et al. “Breast cancer histopathological image classification using convolutional neural networks”. In: *2016 international joint conference on neural networks (IJCNN)*. IEEE. 2016, pp. 2560–2567.

[Sta+16] Daniel Stab et al. “ECG Triggering in Ultra-High Field Cardiovascular MRI”. In: *Tomography* 2.3 (2016), p. 167.

[Ste+04] Robin Steel et al. “Origins of the edge shadowing artefact in medical ultrasound imaging”. In: *Ultrasound in medicine & biology* 30.9 (2004), pp. 1153–1162.

[Sud+17] Carole H Sudre et al. "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations". In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248.

[Sut+13] Ilya Sutskever et al. "On the importance of initialization and momentum in deep learning". In: *International conference on machine learning*. 2013, pp. 1139–1147.

[SZ14] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[Sza04] Thomas L Szabo. *Diagnostic ultrasound imaging: inside out*. Academic Press, 2004.

[Sze+15] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[TBK17] Kerem Can Tezcan, Christian F. Baumgartner, and Ender Konukoglu. "MR image reconstruction using the learned data distribution as prior". In: *CoRR* abs/1711.11386 (2017). arXiv: 1711.11386. URL: http://arxiv.org/abs/1711.11386.

[TBP03] Jeffrey Tsao, Peter Boesiger, and Klaas P. Pruessman. "k-t BLAST and k-t SENSE: Dynamic MRI with high frame rate exploiting spatiotemporal correlations". In: *Magnetic Resonance in Medicine* 50.5 (2003), pp. 1031–1042.

[Tea11] National Lung Screening Trial Research Team. "Reduced lung-cancer mortality with low-dose computed tomographic screening". In: *New England Journal of Medicine* 365.5 (2011), pp. 395–409.

[Uec+14] Martin Uecker et al. "ESPIRiT - an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA". In: *Magnetic resonance in medicine* 71.3 (2014), pp. 990–1001.

[UVL16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. "Instance normalization: The missing ingredient for fast stylization". In: *arXiv preprint arXiv:1607.08022* (2016).

[Vap13] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[Vas+11] SS Vasanawala et al. "Practical parallel imaging compressed sensing MRI: Summary of two years of experience in accelerating body MRI of pediatric patients". In: *2011 ieee international symposium on biomedical imaging: From nano to macro*. IEEE. 2011, pp. 1039–1043.

[Vas+17] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6000–6010.

[Vel+17] Petar Velickovic et al. "Graph Attention Networks". In: *arXiv preprint arXiv:1710.10903* (2017).

[Wag+12] Noam Wagner et al. "Compressed beamforming applied to B-mode ultrasound imaging". In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2012, pp. 1080–1083.

[Wan+04] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[Wan+16] Shanshan Wang et al. "Accelerating magnetic resonance imaging via deep learning". In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2016, pp. 514–517.

[Wan+17a] Fei Wang et al. "Residual attention network for image classification". In: *arXiv preprint arXiv:1704.06904* (2017).

[Wan+17b] Shanshan Wang et al. "1D Partial Fourier Parallel MR imaging with deep convolutional neural network". In: *ISMRM 25th Annual Meeting and Exhibition*. Vol. 47. 6. 2017, pp. 2016–2017.

[Wan+17c] Xiaolong Wang et al. "Non-local Neural Networks". In: *arXiv preprint arXiv:1711.07971* (2017).

[Wan+18] Xiaosong Wang et al. "TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays". In: *CoRR* abs/1801.04334 (2018). arXiv: 1801.04334. URL: http://arxiv.org/abs/1801.04334.

[WB16] Tien Yin Wong and Neil M Bressler. "Artificial intelligence with deep learning technology looks into diabetic retinopathy screening". In: *Jama* 316.22 (2016), pp. 2366–2367.

[WH18] Yuxin Wu and Kaiming He. "Group normalization". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 3–19.

[Wol+13] Robin Wolz et al. "Automated abdominal multi-organ segmentation with subject-specific atlas generation". In: *IEEE transactions on medical imaging* 32.9 (2013), pp. 1723–1730.

[WTF92] Francois Wu, J-L Thomas, and Mathias Fink. "Time reversal of ultrasonic fields. Il. Experimental results". In: *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* 39.5 (1992), pp. 567–578.

[Wur+18] Tobias Wurfl et al. "Deep learning computed tomography: Learning projection-domain weights from image domain in limited angle problems". In: *IEEE transactions on medical imaging* 37.6 (2018), pp. 1454–1463.

[WY13] Ge Wang and Hengyong Yu. "The meaning of interior tomography". In: *Physics in Medicine & Biology* 58.16 (2013), R161.

[XT15] Saining Xie and Zhuowen Tu. "Holistically-nested edge detection". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1395–1403.

[Xu+15] Kelvin Xu et al. "Show, attend and tell: Neural image caption generation with visual attention". In: *International Conference on Machine Learning*. 2015, pp. 2048–2057.

[Yan+15] Zichao Yang et al. "Stacked Attention Networks for Image Question Answering". In: *CoRR* abs/1511.02274 (2015). arXiv: 1511.02274. URL: http://arxiv.org/abs/1511.02274.

[Yan+18] Guang Yang et al. "DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction". In: *IEEE Trans. on Med. Imag.* (2018).

[Yaq+15] Mohammad Yaqub et al. "Guided random forests for identification of key fetal anatomy and image categorization in ultrasound scans". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 687–694.

[YHC18] Jong Chul Ye, Yoseob Han, and Eunju Cha. "Deep convolutional framelets: A general deep learning framework for inverse problems". In: *SIAM Journal on Imaging Sciences* 11.2 (2018), pp. 991–1048.

[YK15] Fisher Yu and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions". In: *arXiv preprint arXiv:1511.07122* (2015).

[YM17] Petros-Pavlos Ypsilantis and Giovanni Montana. "Learning what to look in chest X-rays with a recurrent visual attention model". In: *arXiv preprint arXiv:1701.06452* (2017).

[Yu+17] Qihang Yu et al. "Recurrent Saliency Transformation Network: Incorporating Multi-Stage Visual Cues for Small Organ Segmentation". In: *arXiv preprint arXiv:1709.04518* (2017).

[Zah+18] G Zaharchuk et al. "Deep Learning in Neuroradiology". In: *American Journal of Neuroradiology* (2018).

[ZBF10] Shuo Zhang, Kai Tobias Block, and Jens Frahm. "Magnetic resonance imaging in real time: advances using radial FLASH". In: *Journal of Magnetic Resonance Imaging* 31.1 (2010), pp. 101–109.

[Zbo+18] Jure Zbontar et al. "fastmri: An open dataset and benchmarks for accelerated mri". In: *arXiv preprint arXiv:1811.08839* (2018).

[ZDG19] Martin Zlocha, Qi Dou, and Ben Glocker. "Improving RetinaNet for CT Lesion Detection with Dense Masks from Weak RECIST Labels". In: *arXiv preprint arXiv:1906.02283* (2019).

[Zei12] Matthew D Zeiler. "ADADELTA: an adaptive learning rate method". In: *arXiv preprint arXiv:1212.5701* (2012).

[Zha+16] Hang Zhao et al. "Loss functions for image restoration with neural networks". In: *IEEE Transactions on Computational Imaging* 3.1 (2016), pp. 47–57.

[Zha+17a] Kai Zhang et al. "Learning Deep CNN Denoiser Prior for Image Restoration". In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[Zha+17b] Zizhao Zhang et al. “Mdnet: A semantically and visually interpretable medical image diagnosis network”. In: *CoRR* abs/1707.02485 (2017). arXiv: 1707.02485. URL: http://arxiv.org/abs/1707.02485.

[Zha+17c] Zizhao Zhang et al. “TandemNet: Distilling Knowledge from Medical Images Using Diagnostic Reports as Optional Semantic References”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 320–328.

[Zha+17d] Bo Zhao et al. “A survey on deep learning-based fine-grained object classification and semantic segmentation”. In: *International Journal of Automation and Computing* 14.2 (2017), pp. 119–135.

[Zho+17] Yuyin Zhou et al. “A fixed-point model for pancreas segmentation in abdominal ct scans”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 693–701.

[Zhu+18a] Bo Zhu et al. “Image reconstruction by domain-transform manifold learning”. In: *Nature* 555.7697 (2018), p. 487.

[Zhu+18b] Wentao Zhu et al. “DeepLung: Deep 3D Dual Path Nets for Automated Pulmonary Nodule Detection and Classification”. In: *CoRR* abs/1801.09555 (2018). arXiv: 1801.09555. URL: http://arxiv.org/abs/1801.09555.

[Zog+15] Vasileios Zografos et al. “Hierarchical multi-organ segmentation without registration in 3D abdominal CT images”. In: *International MICCAI Workshop on Medical Computer Vision*. Springer. 2015, pp. 37–46.

[ZSC18] Magauiya Zhussip, Shakarim Soltanayev, and Se Young Chun. “Training deep learning based image denoisers from undersampled measurements without ground truth and without image prior”. In: *arXiv preprint arXiv:1806.00961* (2018).