Reconstruction of Cardiac Magnetic Resonance Images Using Short And Long Axis Slices



Nicolas Basty

Harris Manchester College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2018

Acknowledgements

I would like to express my gratitude to Professor Vicente Grau for his guidance during this project, for believing in me and being encouraging when I decided to change the research focus. I have always enjoyed going into our meetings filled with his *always half full* mindset. A big thank you also to Darryl McClymont for his help in the first year of my research and his feedback on my work after having left the University.

Thank you to my parents for underlining the importance of education and their everlasting support in every way over the course of my studies here and in London. Merci infiniment. Ich bin euch für immer dankbar. Thank you also to Marius, Elsa, and Myriam for being there for me and believing in me. I am grateful to Sarah for being by my side, her patience, and for always bringing a smile to my face. I would not have been able to do it without you.

Also a big thank you to: Kirubin, Ollie, Navin, and Brian. I will cherish memories from 42 for a lifetime. Thank you also to friends and staff from Harris Manchester College, everybody I have met through the CDT, and Wadham College Boat Club. Thank you also to my friends Raphaël, Fritz, and Gilles, also doing PhDs, for keeping in touch over the years and sharing our joys and pains every time we met up.

To the teachers who inspired, motivated, and encouraged me to pursue further scientific education and getting a PhD, Eric Marchand in high school and Kim Parker at Imperial, je dis merci and thank you.

Lastly, I am thankful for the fantastic work environment the University of Oxford has provided, and the EPSRC for funding my DPhil.

Abstract

Cardiac Magnetic Resonance Imaging (MRI) provides high resolution images of the heart, facilitating its functional and anatomical analysis, and is often regarded as the gold standard for diagnosis. Slow imaging times as well as patient and heart motion constraints are making it impractical to image volumes at isotropic high resolution in cardiac MRI. As a result, multi-slice studies, composed of highly anisotropic slices acquired at different orientations, are usually acquired. Most clinical imaging protocols consist in a stack of short axis slices and a few (typically two or three) long axis slices orthogonal to the short axis stack. Post processing methods may be used to reconstruct an isotropic volume. Most of the methods presented in the literature, however, use imaging protocols which are not commonly followed in clinical practice.

In this thesis, we developed three reconstruction methods for cardiac MRI reconstructions using long and short axis images as acquired in clinical practice. First, we showcase our work on a method which does not rely on machine learning or the use of a large database. Our method handles short and long axis data in a contrast independent manner, and is based on regularised least squares and the novel three dimensional directional total variation. We validated the algorithm using pre-clinical data, and showed improvements in reconstructions of in vivo data. We also explored hypothetical protocols for reconstructions, while keeping the number of slices constant. The best results came from balancing short and long axis slices.

We subsequently shifted the research focus to deep learning, allowing a comparison of different methodologies. We used a large publicly available dataset of clinical acquisitions and separated methods into static and dynamic approaches. For the static approach, we present several architectures and show that our networks are able to recover high resolution long axis slices from synthetic data, as well as interpolated short axis stacks. We show that using a conditional Generative Adversarial Network gives the best qualitative and quantitative results, while other models show improved results that are overly smooth. For the dynamic data, we use Long Short-Term Memory. Including temporal context from adjacent frames present in cine data results in enhancement of results compared to the static method. For both methods,

we see qualitative as well as quantitative improvements in peak signal-to-noise ratio and structural similarity over interpolations. Our methods could impact cardiac MRI acquisition and analysis in two ways. In the short term, they could be used to reconstruct missing or additional slices in current acquisitions. In the medium term, they could be used in combination with a fast, optimised acquisition protocol from which slices in any orientation could be obtained at post-processing.

Contents

List of Figures **ix**

1 Introduction **1**
- 1.1 Motivation . . . 2
- 1.2 Contributions . . . 4
- 1.3 Thesis outline . . . 5

2 Background **7**
- 2.1 Cardiac Anatomy and Function . . . 8
- 2.2 Cardiac Disease and Clinical Metrics . . . 10
- 2.3 Modalities used for Cardiac Imaging other than MRI . . . 12
- 2.4 Magnetic Resonance Imaging . . . 15
- 2.5 Cardiac MRI . . . 20

3 Literature review **27**
- 3.1 What is super-resolution, and how is it possible? . . . 28
- 3.2 Super resolution and reconstruction methods in MRI . . . 30
- 3.3 Super resolution in Cardiac MRI . . . 35
- 3.4 Metrics used for quality quantification super resolution . . . 41

4 Reconstruction of Three-Dimensional Cardiac MRI Using Directional Total Variation **45**
- 4.1 Preamble . . . 46
- 4.2 Introduction . . . 49
- 4.3 Data . . . 51
  - 4.3.1 Pre-clinical data (ex vivo) . . . 51
  - 4.3.2 Clinical data (in vivo) . . . 54
- 4.4 Methods . . . 54
  - 4.4.1 Super-resolution algorithm . . . 56
  - 4.4.2 Image acquisition models . . . 59
- 4.5 Results . . . 62
  - 4.5.1 Pre-clinical acquisitions . . . 62
  - 4.5.2 Clinical acquisitions . . . 64
- 4.6 Discussion and Conclusions . . . 65

5 Background on Deep Learning and Literature Review on its application to Super Resolution **71**
5.1 Deep Learning . . . 72
5.2 Convolutional Neural Networks . . . 80
5.2.1 Background . . . 80
5.2.2 The rise in popularity of Convolutional Neural Networks . . . 83
5.2.3 Convolutional Neural Networks for Super Resolution . . . 86
5.2.4 Convolutional Neural Networks for Medical Image Analysis . . . 91
5.3 Generative Adversarial Networks . . . 94
5.4 Recurrence in Deep Learning . . . 99
5.4.1 Recurrent Neural Networks . . . 99
5.4.2 Long Short-Term Memory . . . 100
5.4.3 Applications of Recurrence in Deep Learning to Medical Image Analysis . . . 101

6 Convolutional Neural Networks for Long Axis Slice Reconstruction from Short Axis Stacks **103**
6.1 Introduction . . . 104
6.2 Data . . . 105
6.2.1 Synthetically generated low resolution data . . . 106
6.2.2 Combination of short and long axis slices . . . 107
6.3 Methods . . . 108
6.3.1 SRCNN . . . 110
6.3.2 U-net . . . 110
6.3.3 Conditional GAN . . . 112
6.3.4 Transfer learning . . . 113
6.4 Results . . . 115
6.4.1 Results on synthetic down-sampled long axis data . . . 115
6.4.2 Results on interpolated short axis data . . . 116
6.5 Discussion and Conclusions . . . 120

7 Reconstruction of Long Axis Slice Sequences from Short Axis Stack Sequences Using Convolutional Neural Networks **129**
7.1 Introduction . . . 130
7.2 Materials and Methods . . . 131
7.2.1 Data . . . 133
7.3 Results . . . 136
7.4 Discussion and Conclusion . . . 139

8 Conclusions and further directions **143**

Bibliography **151**

List of Figures

2.1 Cardiac anatomy showing the four chambers, valves and main vessels. Taken from [12]. 9
2.2 Cardiac imaging modalities. Top left: Echocardiography, top right: PET - PET/CT fusion - PET only(from left to right), bottom left: CT, and bottom right: Coronary angiography. 14
2.3 From left to right: Spin alignment in a magnetic field, net magnetisation, a precessing proton. Adapted from [20]. 16
2.4 Spiral precession motion of the magnetisation vector. After being tipped off-axis, M_{xy} slowly decays (T2) while M_z slowly recovers (T1). Adapted from [21]. 17
2.5 Free Induction Decay. Taken from [23]. 19
2.6 Cine images: 1) short axis stack of the first timeframe with 2 long axis views, 2) long axis slice 3) Cut through thick short axis slice stack at the location of the long axis slice. 22
2.7 Short axis slice of a cardiac cine MRI scan with overlay of three long axis slices displaying strong intensity differences between the different orientations. 23
2.8 Standard short and long axis cardiac imaging planes, adapted from [38]. 25
2.9 Different views of a cardiac cine MRI frame. From left to right: 4-chamber view or VLA, 2-chamber view or HLA, mid cavity short axis view. 26

3.1 High resolution grid and a corresponding low resolution grid [41]. . 28
3.2 Pixel shifts in slice selection direction and resulting from a rotation. Adapted from [54]. 34

4.1 Two atoms of coupled low and high resolution 3D dictionaries. . . . 47
4.2 Up-sampling of 3D data using coupled dictionaries. Left: ground truth, middle: low resolution, right: result using dictionaries. The red arrows point edge locations where differences are most noticeable. 47

4.3 Various multi-slice approaches for cardiac MRI coverage. Left: short axis stack in green with a few long axis slices in red (clinical standard), Middle: Stack of short axis in green combined with stack in another orientation in red (as used by e.g. [66, 67, 69, 71]), Right: stack of short axis in green and several long axis slices in red as proposed later in this chapter. 50
4.4 Left: Cut through a high resolution pre-clinical data volume, right: 2D low resolution short axis slice stack and one long axis slice of the same excised rat heart. 52
4.5 Left: cut through 3D view of 11 2D synthetically generated short axis slices resampled into 3D space. The blue plane is aligned with the short axis view while the red and green planes are aligned with long axis views. Right: cut through 3D view of 11 synthetically generated 2D long axis slices resampled into 3D space. The long axis slices have noticeably different contrast than the short axis slices, the myocardium is darker and has lower average intensity values. Adapted from [9]. 53
4.6 6 short axis and 6 long axis slices synthetically generated from HVSMR dataset, patient 17. The artificial intensity difference between short and long axis images is clearly noticeable. The result for this particular reconstruction of multi-slice data is shown in Fig. 4.11. 55
4.7 Top row shows k-space corresponding to the image space data of the bottom row. From left to right: High resolution acquisition, low resolution acquisition, synthetic low resolution acquisition. 61
4.8 Result using 11 short axis and 11 long axis slices.The top row shows reconstructions and the bottom row segmentations achieved by thresholding. From left to right: nearest neighbour interpolation, trilinear interpolation (both interpolations using short axis data only), proposed result, ground truth. Adapted from [9]. 62
4.9 Result using 6 short axis and 6 long axis slices. The top row shows reconstructions and the bottom row segmentations achieved by thresholding. From left to right: nearest neighbour interpolation, trilinear interpolation (both interpolations using short axis data only), proposed result, ground truth. Adapted from [9]. 63
4.10 Result using 6 short axis and 6 long axis slices with the bottom short axis slice missing the apex. 3D slice display. From left to right: nearest neighbour interpolation of short axis slices, our result, ground truth. 64

4.11 Result on synthetic clinical data, using 6 short axis and 6 long axis slices. From left to right: 3D views of nearest neighbour interpolation of the short axis images, the proposed result, and the ground truth volume. This is patient number 17 from the HVSMR dataset. . . . 65
4.12 2D cuts through volumes of synthetic clinical data, using 6 short axis and 6 long axis slices. From left to right: trilinear interpolation (initialisation using short axis data only), proposed result, ground truth. This is patient number 18 from the HVSMR dataset. 66

5.1 A neural network with two hidden layers. 73
5.2 Typical activation functions: sigmoid, hyperbolic tangent, ReLU. . 74
5.3 A visualisation of a two successive 3×3 convolution operations of stride 1 followed by a 2×2 pooling operation. 82
5.4 Architecture of SRCNN, showing patch extraction and reconstruction. The lightweight architecture achieves high quality results at almost real time speeds for up-sampling images by a factor of $\times 3$ [110]. . . 88
5.5 The U-net architecture, taken from [125]. 92
5.6 Top: a Generative Adversarial Network. Bottom: A conditional Generative Adversarial Network for image-to-image translation, both adapted from [134]. 96
5.7 A RNN loop on the left and in its unfolded form on the right, adapted from [144]. A box in the figure on the right hand side is a hidden layer at a given time step, t. 99

6.1 The top row shows 4 examples of synthetic low resolution long axis data, showing two horizontal and two vertical long axis slices. The bottom row shows the corresponding ground truth acquisitions. . . 106
6.2 Left: DICOM data in its original orientation. Right: the same images after applying rotation. 108
6.3 Extracting a long axis slice through the short axis stack. 1. View of short axis stack (blue) and long axis slices (pink) in 3D space. 2. High resolution long axis 4-chamber view. 3. View of the *approximately* equivalent position through the short axis stack. Red arrows are pointing to slice misalignment. 109
6.4 Overlap resulting from transposed convolution. Taken from an illustrative animation on `https://distill.pub/2016/deconv-checkerboard/`. 112
6.5 Interpolations and results for four long axis views, using $SRCNN_{LA}$, the two synthetic U-net models, $U-net_{LA3}$ and $U-net_{LA4}$, $LAGAN$, as well as the ground truth. The figure is flipped to landscape orientation for larger display. 117

6.6 Results on synthetic data of a 2-chamber long axis view. Top row from left to right: nearest neighbour interpolation, bilinear interpolation, ground truth. Bottom row from left to right: $SRCNN_{LA}$, $U-net_{LA3}$, LAGAN. 118

6.7 Interpolation and results for one long axis, using short axis interpolation in training, for SRCNN as well as a U-net model. Top left: nearest neighbour interpolation, top right: linear interpolation, bottom left: $SRCNN_{SA}$, bottom middle: $U-net_{SA4}$, bottom right: ground truth. 119

6.8 Result using real short axis data as input, for SALAGAN. From left to right: short axis input, result using SALAGAN, ground truth. Top: 2-chamber view, bottom: 4-chamber view. 120

6.9 Result using real short axis data as input, for SALAGAN. From left to right: short axis input, result using SALAGAN, and ground truth. 120

6.10 3D views of the two original long axis slices (blue) and a 3rd additionally reconstructed slice (pink), and a different view (bottom) showing intersection of the myocardium of the left ventricle. The orange and blue lines indicate where the planes are intersecting, the profiles along those intersections are shown in Fig. 6.11. 121

6.11 Profiles of the intersections of the true long axis slices with the additional reconstructed long axis slice shown in Fig. 6.10. The orange plot represents the profile along the 3rd long axis slice at the location where it overlaps with the true slices, whereas the blue shows the profile along the ground truth. An arrow is pointing to a dip in both plots, which corresponds to the myocardium. 122

6.12 Result of the original SRCNN architecture trained with combined short and long axis data. 123

6.13 Interpolation, result of $SRCNN_{LA}$, $SRCNN_{LA}$ with residual learning, ground truth. 123

6.14 Discriminator (blue) and L1 (orange) losses for the SALAGAN, trained on short axis data. 125

7.1 Representations of image size scale among the different levels of a U-net. A 128×128 image and its four subsequent down-sampled versions. Each following image has an isotropic decimation factor of 2, resulting in sizes 64×64, 32×32, 16×16, and 8×8. 132

7.2 Network architecture. A network inspired by the U-net with four convolutional LSTM layers at the top of the network aiming to enhance high frequency features by taking into account the dynamic aspect of the cine sequence on the original image size and after the first max pooling operation which decreases the image size by a factor of two. Filter numbers are indicated on the diagram on top of the blue bars after every convolution. Arrows represent the different operations. Height, width, and time sizes are shown for every level. 134

7.3 Result on the first 5 frames of a HLA view cine sequence. The top row shows the low resolution interpolated input, the second row shows the result given by the static U-net, the third the result given by network including LSTM layers, and the bottom row shows the high resolution ground truth. This proposed enhanced 4-chamber sequence has a PSNR of 25.28dB and a SSIM of 0.82 while the static U-net gives a PSNR of 23.83dB and a SSIM of 0.79 and the interpolated sequence a PSNR of 22.15dB and a SSIM of 0.73. . . . 137

7.4 Result on the 1st, 8th, 15th, 22nd, and 30th frames of a VLA view cine sequence. The top row shows shows the low resolution interpolated input, the second row shows the result given by the static U-net, the third the result given by network including LSTM layers, and the bottom row shows the high resolution ground truth. This proposed enhanced 2-chamber sequence has a PSNR of 27.55dB and a SSIM of 0.83 while the static U-net gives a PSNR of 25.69dB and a SSIM of 0.78 and the interpolated sequence a PSNR of 22.23dB and a SSIM of 0.71. 138

7.5 Top: End systole, when the ventricles are at the end of contraction and display the lowest blood volume of the sequence, bottom: end diastole, when the ventricles are at their maximum volumes and about to contract. Red boxes represent zoomed regions. Red dots are present at the same locations as reference points at the end of the myocardium to display the shortening. 141

7.6 Top: interpolated short axis, 2nd row: static U-net result, result on LSTM model, ground truth. 141

1 Introduction

This chapter acts as an introduction the thesis, lays out the motivation and contributions, before giving an outline to the rest of the thesis.

1.1 Motivation

In the United Kingdom, there are over 300,000 deaths related to cardiac diseases every year [1]. While absolute numbers of deaths are decreasing, interventions and numbers of people affected are increasing and cardiovascular diseases still remain the biggest killer in the United Kingdom [2]. Factors affecting this are mainly the ageing of the population as well as others, such as obesity and smoking. These factors are driving the development of novel technology for more accurate diagnosis, preventive monitoring of the population, interventions, and follow-up examinations to monitor patients who have suffered a cardiovascular event. While monitoring and diagnosis of cardiac function can be done via a multitude of methods, the imaging aspect is the focus of this thesis. Within cardiac imaging, a number of modalities can be used for functional and structural assessment. Echocardiography, or cardiac ultrasound imaging, is a non-ionising modality allowing for 2D and 3D real time imaging and is ideal for routine checks as it is cheap and easily accessible, however it suffers from limited resolution, strong artifacts and is highly user dependent. Positron Emission Tomography is an ionising modality which can reveal useful functional information but is often used in conjunction with another modality able to complement with structural information. Computed Tomography is cheaper compared to other modalities capable of giving high resolution 2D and 3D images. CT works by reconstructing several X-ray views into one image and delivers a fairly high radiation dose at each visit [3]. Magnetic Resonance Imaging (MRI) is the modality that produces the best resolution for soft tissue and inter-study reproducibility when compared to echocardiography [4, 5], it is also non-ionising and is hence often taken as the gold standard for imaging. Cardiac magnetic resonance imaging is cardiac specific MRI. It provides both structural and functional information including accurate characterisation of main chambers, vessels and scar tissue as well as the quantification of wall motion, strain and cellular microstructure via diffusion measurements [6].

MRI and especially cardiac MRI however, do not come without any drawbacks. Long imaging times and motion, caused by the patient as well as cardiac motion and breathing, are two of the factors that make MR imaging difficult. These factors have motivated research dedicated to minimising the time a patient is required to spend in the scanner through optimising sequence sampling strategies at the k-space acquisition stage, a method called compressed sensing, and post-processing methods which are often called super resolution. In MRI, compressed sensing stands for sparse sampling of the k-space exclusively, sparse sampling of the image space such as in 2D cardiac MRI is not commonly referred to as compressed sensing.

The subject of this thesis deals with such post-processing methods. We explore a more traditional least squares regularisation method before moving to deep learning approaches, that have been prevalent in current research, on static frames and finally the full sequence. The thesis therefore deals with the comparison of traditional against newly proposed techniques, as well as static against dynamic approaches. Some of the current research deals with deep learning methods which provide direct analysis of medical images [7, 8]. We believe that in a real clinical setting, clinicians will want to see intermediate steps instead of being presented with a simple number, resulting from a black box, for a patient outcome, lung or heart segmentation of 2D images or volumes etc. Therefore we believe it makes sense to decompose the black box into smaller elements, first image enhancement, then segmentation (for example).

The way cardiac MRI is acquired in clinical practice nowadays requires about 20 minutes of scanner time per patient. The aim of our work is to show that using anisotropic 2D multi-slice cardiac MRI data as acquired in clinical practice, it is possible to produce 3D reconstructions, as well as to reconstruct high resolution long axis slices from only using short axis data. The clinical implications of this would be enabling more accurate and reproducible image analysis for diagnosis and monitoring by having isotropic 3D reconstructions, as well as diminishing scanner time required per patient by being able to produce high resolution long axis slices

from exclusively using short axis data and eliminating the long axis acquisitions. In engineering, success is often defined by higher performance determined by metrics. Showing performance enhancements using such metrics can be a way to validate methods, and be used as a proof of concept for further clinical applications when for example fewer data points are required for the same result, which would save time in the data acquisition.

1.2 Contributions

The contributions in this thesis are listed here, grouped into three categories making up each of the analysis chapters.

1. A method for reconstruction of three dimensional cardiac MRI using directional total variation as a regulariser. Directional total variation was previously only used in 2D. For this method we introduce a novel image acquisition model for pre-clinical images based on truncating the k-space. We also argue that the current clinical protocol with a stack of short axis and only few long axis images is far from ideal, and show that using a balanced number of short and long axis slices improves reconstructions. These contributions are explored in chapter 4.

2. A deep learning approach for super resolution of static MRI frames. The data pre-processing for short axis relies on an acquisition model different from the previous acquisition model, which instead of truncating the k-space, uses gaussian averaging. It can be used to acquire any additional long axis view through an interpolated short axis stack. In addition to that, the use of transfer learning on generative adversarial networks on single shot learning is explored. These contributions make up chapter 6 and some of the data pre-processing is used in chapter 7 as well.

3. A deep learning approach for super resolution of dynamic MRI frames. This method combines long short-term memory layers with the U-net architecture using a novel approach to the selection of combination layers. These contributions make up chapter 7.

1.3 Thesis outline

The remainder of this thesis is organised as presented in this outline.

Chapter 2 introduces the thesis further by giving the reader background information about cardiac anatomy, function and pathophysiology before going into cardiac diagnostic methods with a focus on MRI physics as well as their application to clinical cardiac MRI.

Chapter 3 presents the reader with a literature review of classical methods used at the acquisition stage and at pre-processing to address the issues associated with cardiac MRI outlined in the motivation.

Chapter 4 presents a new approach to reconstruction of three-dimensional cardiac MRI via regularisation. We present the first use of three-dimensional directional total variation. We use an acquisition model working in three-dimensional k-space. We validate the algorithm using high resolution pre-clinical rat hearts, some specifically acquired for this project. Further in this chapter, we use the method to reconstruct 3D volumes by applying it to synthetic low resolution 2D data generated from a high resolution 3D data set, which is used as ground truth. We presented part of this work dealing with validation on pre-clinical data as an oral presentation at the workshop on Reconstruction and Analysis of Moving Bodies (RAMBO) at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017 and it is published in the proceedings [9].

Chapter 5 presents the reader with background on deep learning methods and a

literature review on the application of deep learning methods to super resolution problems as well as medical image processing to support the following two chapters on deep learning approaches.

In **Chapter 6** we present the application of benchmark techniques to the problem of synthetic anisotropic 2D cardiac MRI data before moving on to our use of generative adversarial networks on real data affected by issues mentioned before. The final part of this chapter is dedicated to our work using data from a single heart applied to fine tune a network via transfer learning. We presented a poster on some of the work from this chapter at the International Symposium on Biomedical Imaging (ISBI) 2018 in Washington DC, USA. A paper with the tentative title "Reconstructing long axis slices from short axis stacks in cardiac MRI using deep learning" dealing with the work of this chapter is in the planning stages.

Chapter 7 presents our work on super resolution which is taking advantage of the temporal aspect of cine MRI. The dynamic aspect of medical images, and especially cardiac MRI is something that is often ignored in the literature. We incorporate long short-term memory layers into a U-net, a well established network architecture, at high frequency levels for the first time. We presented part of the work in this chapter as an oral presentation at the RAMBO workshop of the MICCAI 2018 conference in Grenada, Spain, and it is published in the proceedings [10].

Chapter 8 provides a final conclusion to our work and indicates possible further directions based on the results presented in this thesis. In addition to the possibility for further post processing research using deep learning, this thesis opens up a discussion to optimise imaging protocols for reconstructions based on the three analysis chapters.

2
Background

This chapter further introduces the subject matter of the thesis by giving the reader background information about cardiac anatomy, function and pathophysiology before going into cardiac diagnostic methods with a focus on MRI physics as well as background on cardiac MRI.

2.1 Cardiac Anatomy and Function

The function of the heart is to pump blood through the cardiovascular system in order to transport oxygen and nutrients to other organs, as well as removing metabolic waste products such as carbon dioxide. The heart is a dual pump system containing four chambers, the right and left atria and the right and left ventricles. A diagram of the heart anatomy is shown in Fig. 2.1. The heart is partitioned into the right side and the left side, each containing an atrium and a ventricle, working in sequence during the cardiac cycle. The two sides are separated by the cardiac septum, which prevents contact of of deoxygenated with oxygenated blood. Blood enters the left atrium via the pulmonary veins and right atrium via the inferior and the superior vena cava. The right side drives the pulmonary circulation, bringing deoxygenated blood through the lungs to be re-oxygenated before entering the left side which distributes the blood into the systemic circulation. The pumping motion is triggered by electrical stimuli in pacemaker cells emitting a signal in the sinoatrial node of the right atrium which spreads across the tissue and is conducted through to the atrioventricular note to the ventricles, which then contract. At rest, the heart beats at a steady rate which varies between individuals. There are two main phases in the cardiac cycle: diastole and systole. Valve mechanics, allowing or preventing blood flow, are pressure driven.

During systole, the ventricles contract, causing the pressure inside the chambers to increase. Once a threshold pressure is exceeded, the aortic and pulmonary valves open and let the oxygenated blood from the left ventricle flow through the aorta into the systemic circulation while the deoxygenated blood in the right ventricle flows through the pulmonary artery towards the lungs to be re-oxygenated. During diastole, the atria and ventricles relax causing the pressure to drop. Atrioventricular valves open and the right ventricle and atrium are filled with deoxygenated blood while the left ventricle and atrium are filled with oxygenated blood. The heart itself also needs blood to function, this need is met by the coronary circulation system which branches out of the aorta at its base [11].

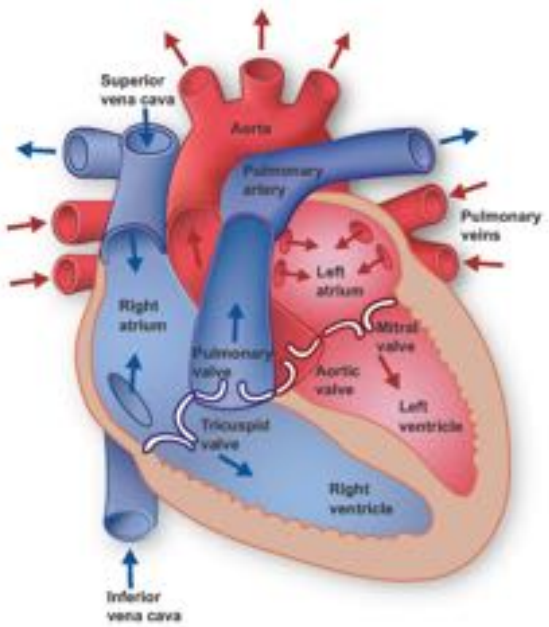


Figure 2.1: Cardiac anatomy showing the four chambers, valves and main vessels. Taken from [12].

The heart and the roots of the great vessels are lodged between the lungs in the thoracic cavity, and kept in place by the pericardium in proximity of the sternum and diaphragm. The pericardium is a double walled membrane which provides a smooth environment for the myocardial surfaces during the cardiac cycle. The orientation of the heart is approximately in line with an axis connecting the right shoulder and the left hip. The muscular tissue making up the walls of the cardiac muscle is composed of three layers: the epicardium, an additional protective layer forming the outer wall, the myocardium, and the endocardium which is the inner layers of the wall muscle. The myocardium is the thickest of the three layers responsible for the contraction of the heart, mainly composed of cardiomyocytes. The endocardium is made from endothelial cells allowing a low friction surface with the blood [13]

2.2 Cardiac Disease and Clinical Metrics

The above brief introduction to the anatomy and structure of the heart showed that the heart is an organ made from a complex structure involving pressure gradients triggered by electric impulses. There exist various causes for cardiac disorders related to the three aforementioned properties: they can impede the electrical, structural, mechanical, or circulatory properties of the heart. Electric and structural diseases many times have a genetic component, while most cardiovascular diseases can be traced back to behaviour such as an unhealthy diet, lack of exercise or smoking. Electrical defects are the cause of arrhythmia, or abnormal heart rhythms. They can lead to sudden cardiac arrest, which is responsible for a large number of sudden deaths in otherwise healthy young individuals. In the last decade, the incidence of sudden cardiac arrest in England and Wales has more than doubled from 13.5% to 28.1%, and now matches the incidence of ischemic heart disease for cardiac related deaths of young people (1 to 34 years old) [14]. Structural issues of the heart include patent foramen ovale, valvular heart disease and hypertrophic cardiomyopathy. Patent foramen ovale is a mild congenital condition where a small opening remains after birth in the cardiac sternum. Valvular disease affects the ability of the heart to move blood from chamber to chamber efficiently. Hypertrophic cardiomyopathy produces an excess thickening of the myocardium, which can lead to heart failure as it affects the efficacy of the pumping function. Circulatory diseases relate to issues in the cardiovascular system. The majority of circulatory diseases are coronary artery diseases (ischemic heart diseases), describing a lack of oxygen supply to tissue or hypoxia. The main cause of ischemic heart diseases is atherosclerosis, which is the narrowing of arteries happening because of the accumulation of fatty deposits called plaque. The severity of ischemic heart disease is dependent on the percentage of occlusion. It ranges from angina to myocardial cell death, or myocardial infarct, also often called a heart attack. Angina occurs when the blood supply to the heart is temporarily not met, usually caused by obstruction of the coronary arteries. Angina refers to short periods of reduced oxygen supply, and does not induce long term damage to the heart. Myocardial infarctions occur when blood flow through

the arteries is partially or fully obstructed and result in longer periods of hypoxia leading to cell death. Most commonly, myocardial infarcts occur due to a blockage in coronary arteries. If not recognised and treated immediately, it can lead to irreversible damage and scar formation, or be fatal. Long term effects on cardiac tissue depend on the severity and the duration of the ischemic event [15]. Plaque can also lead to the formation of blood clots. Such blood clots, or pieces of plaque itself, can move through the vasculature and trigger blocking of arteries further downstream leading to strokes or transient ischemic attacks, depending where the occlusion occurs.

The heart is one of the vital organs and needs accurate means of diagnosis and monitoring for both structural and functional assessment. It can be imaged using various ways, of which MRI is the gold standard as previously mentioned.

Several quantitative metrics are available for assessing cardiac function including stroke volume, ejection fraction and cardiac output. These can be evaluated from various imaging modalities. Stroke volume quantifies the volume of blood pumped during a cardiac contraction. It is calculated by taking the difference between the largest and smallest left ventricular volumes, at the end of diastole and at the end of systole, respectively. The ejection fraction corresponds to a ratio, where the volume of blood ejected by the heart during a contraction is divided by the end-diastolic volume. It is used as an early indicator for systolic heart failure. Reduced ejection fraction is also a consequence of heart remodelling after a myocardial infarct. Healthy tissue dilates to counter balance the lack of function from infarcted tissue which compromises the function, which is directly measurable by a lower ejection fraction. The cardiac output measures the total output volume of blood expressed as a function of time, by multiplying the stroke volume by the heart rate. The cardiac output allows clinicians to identify wether the heart is able to supply the amount of blood meeting the demands of the rest of body, which is the main function of the heart. Other means of evaluation of cardiac function include wall motion tracking

and valvular function. All of those indicators of cardiac function can be derived from utilising image and video data obtained by cardiac imaging. The next section will deal with cardiac imaging. This thesis deals with data from MRI exclusively, and thus the following section will mainly focus on MRI. There exist several different non-invasive imaging modalities, and depending to the problem at hand, clinicians have to decide which modality to use. The decision is made keeping in mind if they want to obtain structural or functional information, and the degree of image quality while keeping the ionisation dose to a minimum or ideally no radiation at all.

2.3 Modalities used for Cardiac Imaging other than MRI

Cardiac imaging is used to obtain insight about cardiac anatomy and function. Depending on urgency, as well as the availability of methods, clinicians make a decision on which imaging modality to use. In the following, a very brief view of four major modalities is given before a more in depth view of MRI.

Echocardiography

Echocardiography is the name given to cardiac-specific ultrasound imaging. Ultrasound pulses are sent through tissue, and images are formed by converting the measured reflected signal into image intensities. Imaging can be done in 2D, 3D and also using Doppler ultrasound. The latter uses frequency or the phase shift to determine the speed and direction of blood flow. There is also tissue Doppler, which detects speed and direction of tissue movement. Echocardiography is non-invasive and non-ionising and is routinely used for cardiac examinations. It allows for rapid examinations, and offers a broad range of applications including valvular function. Depending on the desired orientation, sonographers can use different acoustic windows to achieve various views of the heart. Image quality is highly dependent on the level of experience of the sonographer, which makes echocardiography prone to inter-user variability and may affect reproducibility. An

example of a transthoracic 2D apical 4-chamber view displaying valves, atria and ventricles can be seen in the top left of Fig. 2.2 [16].

Computed Tomography

Computed Tomography (CT) is a non-invasive ionising imaging modality based on x-ray technology. X-rays, discovered at the end of the 19$^{\text{th}}$century, gave rise to radiography which was the first non-invasive medical imaging technique. Radiography images are formed by collecting x-ray radiation passing through the body, where various tissues and bones attenuate the radiation, which becomes visible when the radiation is collected on the other side of the patient, on film or more recently image intensifiers based on charged coupled devices. During a CT exam, numerous x-ray beam projections are obtained at different angles. The back projection technique is used on the raw data from those projections to reconstruct an image, often in isotropic three dimensional resolution. The higher the number of projections, the better the quality, but with it comes an increase in administered dose. The two main applications of CT to cardiac imaging are coronary CT angiography and coronary CT calcium scans. The former requires administration of a contrast agent and is used to assess the degree of atherosclerosis, the latter does not require the use of contrast agents and is used to detect calcium deposits that can be indicators of increased susceptibility to myocardial infarcts [17]. An example of a CT calcium scan is shown in the bottom left of Fig. 2.2 [18]. The left hand side shows evident calcification, the right hand side shows the same scene where the calcium is highlighted in pink, using an intensity threshold.

Positron Emission Tomography

Positron Emission Tomography (PET) is an imaging modality based on nuclear medicine. Radioactive tracers are injected, and their decay releases positrons which annihilate with electrons, a process which produces photons. The photons are collected by a ring of radiation detectors and used to form the image. Because the

tracers are absorbed depending on tissue health and function, PET is able to give metabolic information about the heart. Some of the uses of PET are to quantify myocardial perfusion and myocardial blood flow at rest and during stress. PET is often used in conjunction with CT or MRI, to fuse functional information from PET and put it into structural context, since PET gives poor structural information [18]. An example of stress (top) and rest (bottom) myocardial perfusion PET imaging can be seen in the top right of Fig. 2.2. The left hand side shows the CT providing structural information, while the right hand side shows the functional information obtained from PET, and the centre shows the fused images.

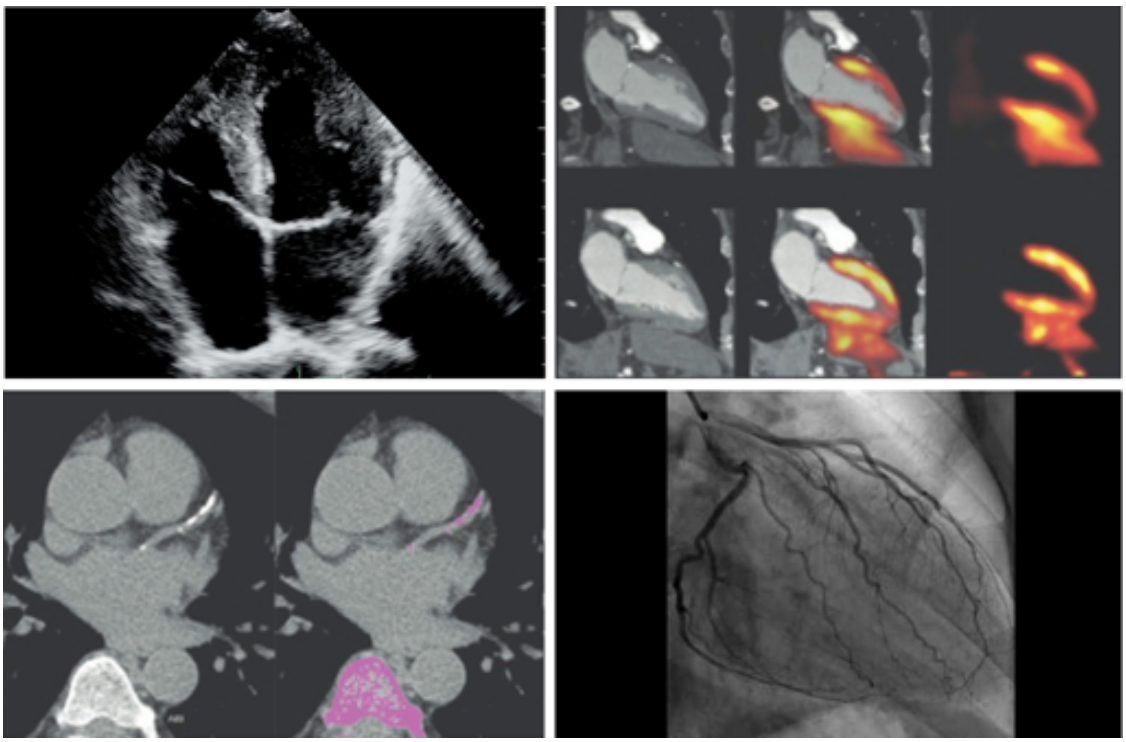


Figure 2.2: Cardiac imaging modalities. Top left: Echocardiography, top right: PET - PET/CT fusion - PET only(from left to right), bottom left: CT, and bottom right: Coronary angiography.

X-ray angiography

X-ray angiography is a standard technique for the imaging of coronary arteries. It is used to determine the presence and degree of obstructive coronary artery disease when non-invasive testing is not conclusive, to assess the feasibility and appropriateness of therapy, and for the assessment of treatment results and to monitor the progression or regression of atherosclerosis. The decision for choosing this modality must be made on expected benefits balanced against the risks of the technique, according to governing bodies such as the American Heart Association [19]. It is often used in acute settings and often involves catheterisation, making it a minimally invasive procedure. Iodine-based contrast agents are regularly used. In conjunction with cine fluoroscopy, a dynamic imaging modality based on x-rays, it allows cardiologists to receive real time feedback during procedures such as stent placement operations. As a modality involving x-rays, just like CT, it is ionising to the patient, but also to the practitioners when used in acute settings [17]. An example of x-ray angiography is shown in the bottom right of Fig. 2.2, the vessel tree is clearly discernible as a result of the injection of a contrast agent.

2.4 Magnetic Resonance Imaging

MRI is a non-ionising and non-invasive imaging method allowing for particularly good soft tissue contrast. It creates images by measuring the response of atoms to radio-frequency (RF) pulses when placed within a strong magnetic field. It allows for distinction between viable and dead tissue, characterisation of perfusion and motion, amongst other soft tissue observations, making it a good modality for imaging and assessing the myocardium. It is based on the principles of NMR, nuclear magnetic resonance. NMR is a spectroscopic technique used for giving insight of microscopic physical and chemical information about the nuclei of a molecule.

The nucleus of an atom can contain positively charged protons and neutrally charged neutrons. Spin is a quantum mechanical property of elementary particles,

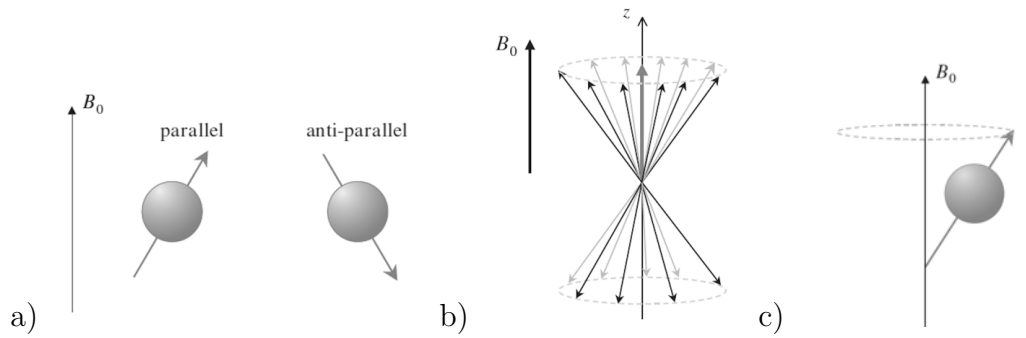


Figure 2.3: From left to right: Spin alignment in a magnetic field, net magnetisation, a precessing proton. Adapted from [20].

that can be visualised as spheres spinning and generating a weak electric current which generates a magnetic field, aligned with the so-called spin vector. Spins in neutrons and protons have opposite spins and magnetic fields. NMR and MRI are based on this magnetic property of atoms, so nuclei that do not have a net magnetisation because the sum of the magnetisations of their nucleons is zero are not useful. The nuclei of Hydrogen atoms are made up of a single proton, have no neutrons counterbalancing their spin, and therefore have a net magnetisation and are very useful in MRI for this reason. In the human body, hydrogen atoms in water are the most abundant atoms.

When an object or patient are placed in a strong magnetic field B_0 (common clinical magnets have B_0 of 1.5 to 3T (Tesla), for comparison: the earth's magnetic field is roughly 0.05mT which is enough to align the needle in a compass), the spins of the protons in hydrogen atoms align with the magnetic field of the scanner in a parallel or anti-parallel manner, as shown in Fig. 2.3 a). This generates a net magnetisation vector M, which has zero magnitude M_{xy} perpendicular to $\mathbf{B_0}$, shown in Fig. 2.3 b). When placed in a static magnetic field, particles with non-zero spin precess about the direction of the magnetic field, a process called gyromagnetic precession. The precession frequency of a nucleus in a magnetic field is called the Larmor frequency ω, and is proportional to the magnitude of the magnetic field

and γ_0, the gyromagnetic ratio of the atom: $\omega = \gamma_0 B$ [21]. A proton precessing about B_0 can be seen in Fig. 2.3 c).

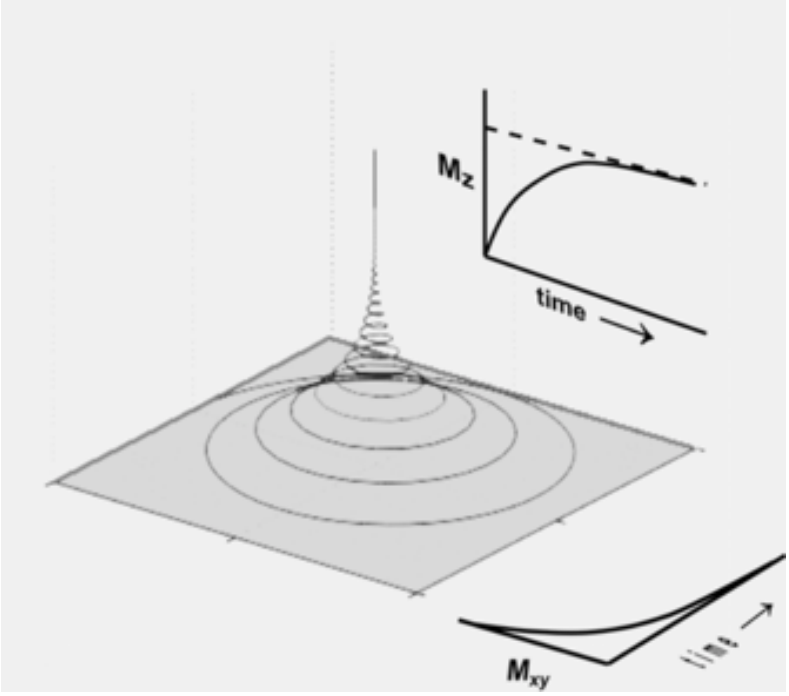


Figure 2.4: Spiral precession motion of the magnetisation vector. After being tipped off-axis, M_{xy} slowly decays (T2) while M_z slowly recovers (T1). Adapted from [21].

The net magnetisation vector M can be tipped off axis by applying RF pulses at the resonant frequency of the atoms in question, in this case hydrogens in water, and transferring energy to the atoms one wants to excite via the resonance principle. During the application of the RF pulse, the in-plane magnetisation M_z decreases while the out-of-plane magnetisation M_{xy} increases. The amount of perturbation of M is also called the flip angle, measuring the angle spanning between M and B_0. The flip angle depends on the magnitude and duration of the RF pulse, and is often chosen to be 90° to fully eliminate M_z. After applying a short RF pulse, the protons lose the excess energy by precessing back in a spiral around the direction of field of the magnet. Fig. 2.4 shows the spiral precession as well as the recovery

of M_z and decay of M_{xy}. This results in the net magnetisation returning to its initial orientation and magnitude, by two processes called spin-lattice and spin-spin relaxation. Spin-lattice relaxation described the recovery of M_z, it is also called the longitudinal or T1 relaxation. T1 is the time needed for M_z to recover $(1 - e^{-1})$ or 63% of M_0, the equilibrium magnitude. The surrounding environments (the lattice) play a role in the rate at which energy is lost by the spins. For example, the motion of lipid molecules is conductive to energy transfer which makes the T1 of lipids short. The T1 of water is longer, as its molecules have higher molecular motion. Spin-spin relaxation is also known as the transverse relaxation or T2 relaxation, and describes the decay of M_{xy}. T2 is the time needed for M_{xy} to decay to e^{-1} or about 37% of M_0. Spin-spin refers to the loss of phase coherence between the spins themselves. Like with T1, the environment affects the times. High water content leads to higher T2 values, because of the increased space between molecules making spin-spin interactions less likely. T1 and T2 relaxations occur simultaneously but T2 is much faster. T2 decay is also affected by magnetic field inhomogeneities, which has led to the introduction of the additional T2* constant. It is described as $\frac{1}{T2*} = \frac{1}{T2} + \frac{1}{2\gamma\nabla B}$. The T2* is always shorter than T2, which accelerates the T2 decay. This can be understood as having the inhomogeneities speeding up the dephasing process and this is why T2* is referred to as the "observed T2". The processes are governed by the Bloch[1] equations: $M_z(t) = M_0(1 - e^{-\frac{t}{T1}})$ and $M_{xy}(t) = M_0(e^{-\frac{t}{T2}})$ [20]. The two equations are plotted against time in Fig. 2.4.

The M_{xy} component of the net magnetisation is at the source of the signal measured in MRI. From the frame of reference of the RF coil, which also acts as a receiver, M_{xy} relaxes whilst spinning around the Z-axis at the Larmor frequency due to the rotation of the magnetic field vector. This results in a decaying sinusoidal function when plotted over time. The decaying rotating magnetic field generates an induced current, and is called the Free Induction Decay (FID). Fig. 2.5 shows the FID. The FID describes the decay of M_{xy} from the RF coil reference frame. The

[1] named after Felix Bloch, Swiss physicist who was awarded a share of the 1952 Nobel Prize for Physics

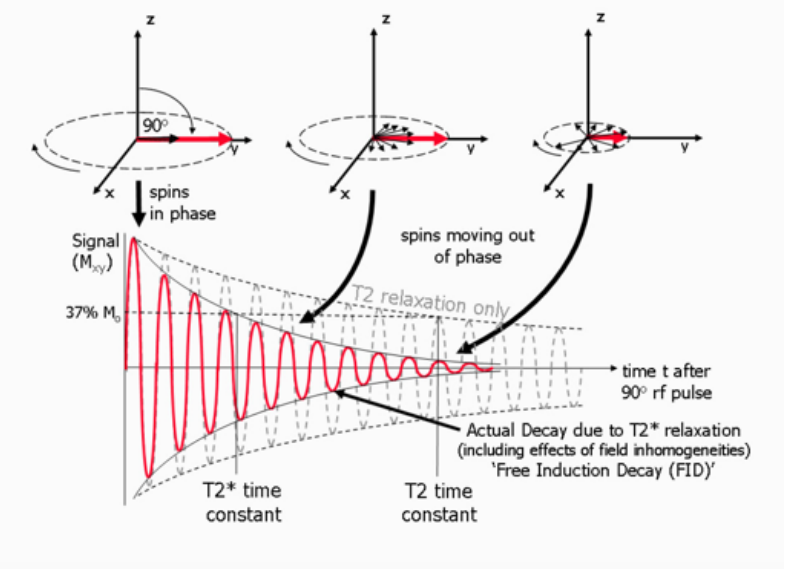


Figure 2.5: Free Induction Decay. Taken from [23].

intensity of signals perpendicular to B_0 will be different from zero as long as the net magnetisation vector is perturbed, and this is the signal that is picked up using coils. Images are obtained by evaluating the locations and intensities from the FID signals. This is done by applying magnetic field gradients, creating varying field strengths depending on the position. Planes are defined by so called "slice selection" pulse, an RF pulse of a given bandwidth which defines the thickness of the slice to be excited. The signals recorded by the coils make up the so called k-space, taking a Fourier transform of the k-space produces an image [22].

There exist various pulse sequences used to generate images, spin echo being one of the most common group of sequences. It uses a second RF pulse called the inversion pulse that inverts the magnetisation vectors to cancel out the dephasing effect. Within spin echo, the three ways of controlling the contrast in MRI acquisitions are called T1, T2, and Proton Density weighting. They are defined by different repetition times (TR) and echo times (TE) in MRI sequences. TR describes the time between successive pulse applications and determines the time allowed for relaxation processes so image contrasts can be manipulated that way. The echo

corresponds to the peak of the FID signal induced in the coil, so the TE is the time taken between the RF pulse and the echo. The T2 relaxation can be fine-tuned by adjusting the TE. T1 weighted imaging gives tissues with short T1, such as lipids, higher intensities while water-filled compartments appear darker on the image. They are achieved by short TR and short TE, to minimise T2 relaxation. T2 weighted images are generated using long TR and long TE, giving a higher signal intensity to tissues with long T2. Fatty tissues have short T2 and water has a long T2, so for long TE fat will appear darker than water. The third is Proton Density, it is achieved with long TR and short TE. Long TR reduces T1 weighting and short TE reduces T2 weighting, leading to an image formed without specific T1 or T2 weighting but rather one that reflects the amount of signal from each location [24].

While MRI works very well for imaging of static organs or soft tissues such as the brain, the constant cardiac, breathing and general patient motions introduce additional problems to acquisitions of the heart due to relatively long image acquisition times.

2.5 Cardiac MRI

Cardiac MRI addresses the issue of cardiac motion by using a different pulse sequence and gating imaging times to electrocardiogram readings, while breathing motions are alleviated by patients being told to hold their breath. Even though there are limitations to Cardiac MRI, given its great accuracy, reproducibility and safety considerations, it is considered as the gold standard measurement for ventricular function parameters such as volumes or ejection fraction [25, 26]. In addition to providing excellent resolution and therefore anatomical information [27], wall motion may be analysed using MR data, further adding to the significance of Cardiac MRI [28–30]. Recent studies also focus on the application to cardiac electrophysiology [31, 32].

The MRI pulse sequence most commonly used in cardiac cine MRI for left ventricular structural and functional analysis is the balanced steady-state free precession (b-SSFP) sequence. The b-SSFP is a gradient echo MRI sequence, these are sequences using only one RF excitation and a gradient. The gradient is used to dephase the spins and add spatial dependence. The echo is triggered using the gradient, hence the name gradient echo, instead of a second pulse like in spin echo sequences. Fast acquisition times are aided by low flip-angles and very short TR. In b-SSFP, TR are about 3ms against about 30ms in spin echo sequences. Fast acquisitions and electrocardiogram gating make b-SSFP ideal for Cardiac MRI. In addition to that, b-SSFP also has excellent signal-to-noise ratio per unit time and T2/T1 contrast, and it does not suffer from excessive signal loss from motion [33]. There also exists a 3D version of the b-SFFP sequence, which allows isotropic acquisitions and to image the full volume, but in turn has worse contrast between blood and the myocardium and is therefore not commonly used in clinical practice.

Cardiac cine MRI is the standard technique for combined cardiac structural and functional imaging of the heart, and also the kind of clinical data available for our study. Multiple images, or frames, covering the cardiac cycle over one beat are acquired in order to capture the cardiac motion. Patients are asked to hold their breath to minimise resulting motion artifacts and misalignment between images. In practice, each signal is acquired twice with the aim to average breathing motion artifacts which further increases acquisition times. In cine MRI, the most widely used pulse sequence is b-SSFP due to its short acquisition times, independence of contrast from blood flow and high contrast to noise ratio [6]. While most of the heart may be covered in the acquisitions, the resolution is highly anisotropic with good in-plane resolution and much worse out of plane resolution. An example of an image stack at one particular timeframe can be seen in Fig. 2.6 1), as well as a long axis acquisition in 2) and a matching cut through the stack represented as thick slices 3). 3) shows how much, or how little of the space is actually acquired in the short axis stack compared to what long axis views cover. The intensity differences

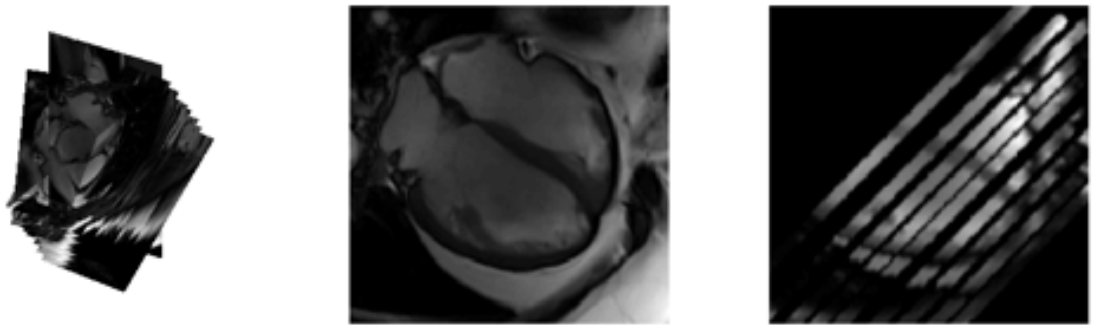


Figure 2.6: Cine images: 1) short axis stack of the first timeframe with 2 long axis views, 2) long axis slice 3) Cut through thick short axis slice stack at the location of the long axis slice.

between long axis and short axis, as well as between the short axis slices themselves are also visible.

There are several issues related to such intensity differences across different MRI acquisitions of the same organ as well as inconsistencies within an image itself. One of the most prominent artifacts are inconsistencies of pixel intensity caused by the so called bias field. It generates artifacts noticeable as faint gradual changes in pixel intensities in the image domain that do not reflect tissue characteristics, but are due to imperfections in the magnetic field of the scanner. Homogeneous objects in the image may display a variation in pixel intensities in the acquired image. Various prospective and retrospective methods of intensity correction for bias field related inhomogeneities exist and are explored in [34]. Another issue is the absence of true meaning of pixel intensities in cine MRI. Unlike CT where every tissue is assigned a CT number on the Houndsfield scale calibrated to water, pixel intensities in cine MRI have no direct meaning or scale. In two different images, pixels belonging to the same tissue, organ, or fluid may have completely different intensities. In addition to that, intensity differences may vary across tissues, so inconsistencies

between images may not be solved by simple scaling. Intensity inconsistencies and differences make analysis of single images such as segmentation difficult, but also add to the difficulty of combination of various views of the same object, something directly related to multi-slice cardiac MRI acquisitions. This problem is highlighted in Fig. 2.7, which shows a cut at a short axis location through a 3D image combining short and long axis scans from a frame of a cardiac cine MRI acquisition. The long axis slices are clearly discernible with the naked eye, highlighting the difficulty of incorporation of different views of clinical multi-slice cardiac MRI data.

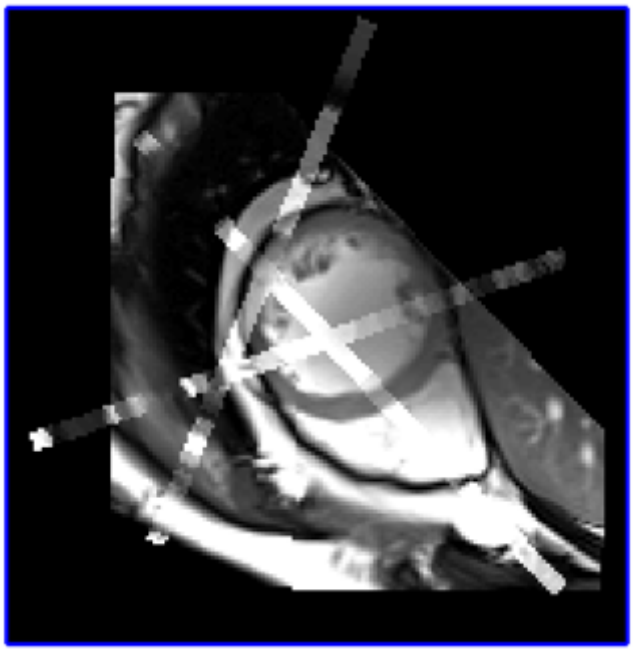


Figure 2.7: Short axis slice of a cardiac cine MRI scan with overlay of three long axis slices displaying strong intensity differences between the different orientations.

Due to the motion and time constraints, high resolution 3D cardiac images are difficult to acquire. Indeed, 3D b-SSFP is an option, but it is not in widespread clinical use. Instead, in a clinical routine cine MRI examination, anisotropic slices are acquired instead of a 3D volume. The common acquisition protocols include a stack of parallel short axis slices as well as a number of long axis slices approximately orthogonal to the short axis stack. The number of slices in the short axis stack

is dependent on the size of the heart and the number of long axis slices is also variable. Standardised protocols including the UK biobank protocol [35,36] include three long axis views: the vertical long axis (VLA, also called 2-chamber view), the horizontal long axis (HLA, also called 4-chamber view), and the left ventricular outflow tract (LVOT) view while others such as the American Heart Association may only include two: the HLA and VLA [37]. Short axis views correspond to slices taken longitudinally to the heart and long axis ones taken transversally. Short axis slices may be taken from the apex to the base and can contain the left ventricle on its own or both the left ventricle and the right ventricle, depending on the location of the slice. These standardised protocols have been set in place for large scale efficient, safe, and reproducible imaging. In the UK Biobank case, the aim of the study is to achieve large scale cardiovascular phenotyping. For this, quantifying metrics relating to left and right ventricular as well as atrial and aortic parameters are being evaluated. The committees in question selected MRI for its safety and reproducibility characteristics and a multi-slice sparse imaging for time concern. The protocol that resulted from the rationale allow for characterisation of those parameters and have been based on clinical practice for the same rationale.

Fig. 2.8 shows standard imaging planes in cardiac MRI. Short axis views are shown in A (basal), B (mid-cavity), and C (apical) while long axis views are shown in D (VLA) and E (HLA). The numbers 1-17 correspond to the segments of coronary arterial territories, as defined by the standardised American Heart Association 17-segment model [38]. Fig. 2.9 shows a the HLA, VLA, and a mid-cavity short axis frame of Cardiac MRI and labels on the locations of each of the two or four chambers.

Some of the main issues associated with cardiac cine MRI are the slice misalignment occurring due to patient and cardiac motion between acquisitions, intensity differences between slices due to flow artifacts and magnetic field inhomogeneities, and sometimes contrast agents, as well the sparsity of the data occasionally resulting in a lack of coverage of the left ventricle by the short axis stack [39].

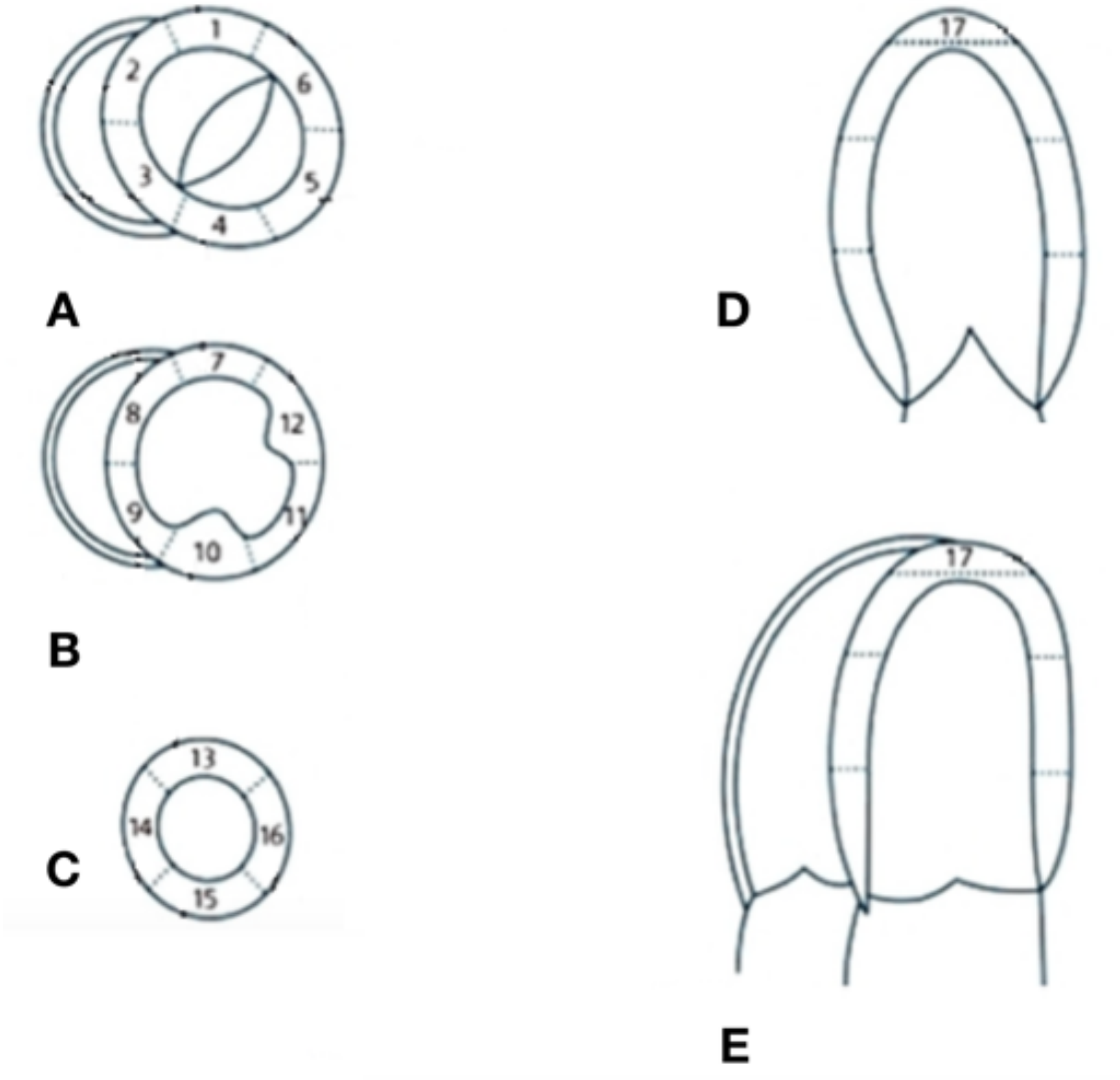


Figure 2.8: Standard short and long axis cardiac imaging planes, adapted from [38].

Recently, the use of cardiac MRI has become more prominent in acute conditions such as myocardial infarctions, as better tools for diagnosis and risk estimation are needed where early interventions may prevent irreversible injuries. A better characterisation of the myocardium shortly after an infarct or as chest pain is setting in could provide additional information to the standard tools such as echocardiography which may be used for evaluating parameters such as valve function but is limited in soft tissue contrast. Especially in an acute setting, reducing the imaging time is of great interest [40]. However, even in non-acute settings, it is preferred to keep time to a minimum, as patients get claustrophobic and may be averse to this modality. In

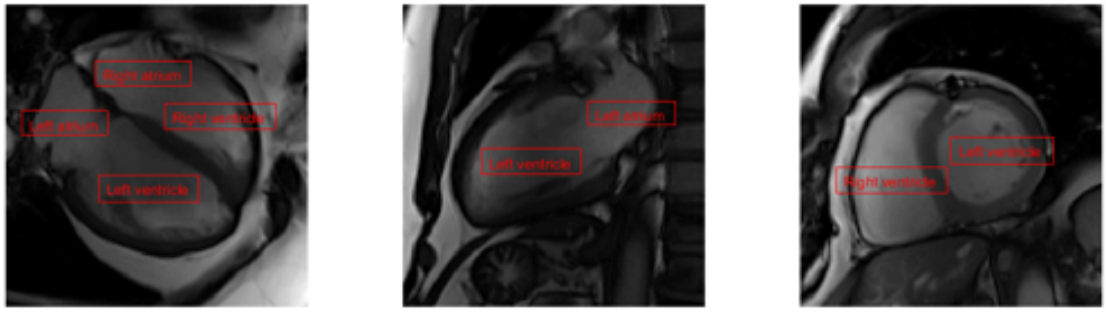


Figure 2.9: Different views of a cardiac cine MRI frame. From left to right: 4-chamber view or VLA, 2-chamber view or HLA, mid cavity short axis view.

addition to this, MRI facilities are highly expensive and require a large amount of space and are hence not readily available all the time depending on the institution.

3
Literature review

The previous chapter presented the reader with knowledge on cardiac anatomy, cardiac imaging, with particular focus on MRI and cardiac MRI. It also laid out the motivation for research for smart k-space acquisition protocols also known as compressed sensing, as well as image reconstruction and super resolution techniques and indeed this thesis which deals with the latter. This chapter reviews literature related to more traditional post-processing methods for reconstruction and super resolution in MRI, mostly based on regularised least squares.

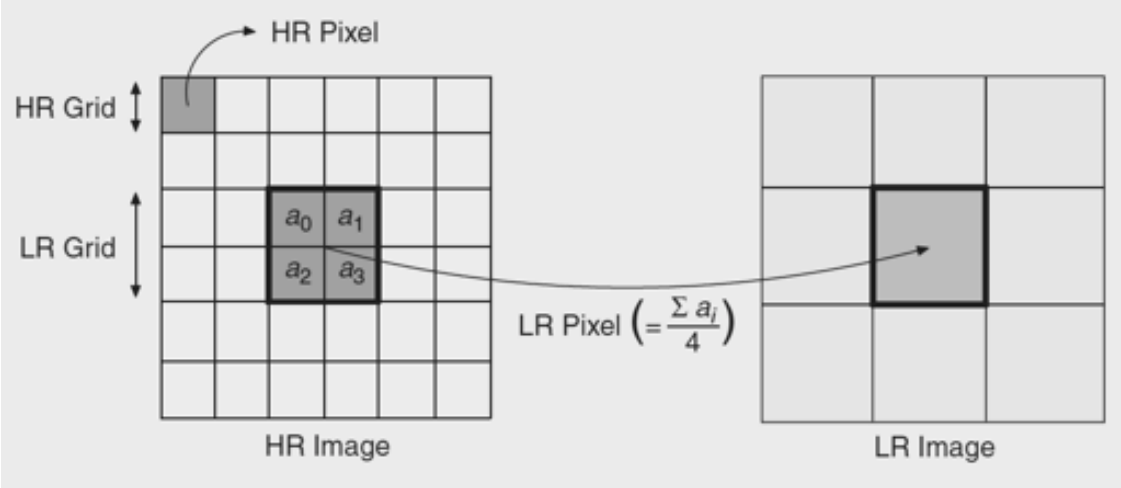


Figure 3.1: High resolution grid and a corresponding low resolution grid [41].

3.1 What is super-resolution, and how is it possible?

Super resolution is a process by which a single image, or a set of images are processed to create an image or a set of images with higher resolution than the original data. The resolution and general quality of images depend on several factors, including the limits of the imaging device and motion blur resulting from motion of the imaging device or the scene being imaged. High resolution is synonymous with high pixel density, meaning that high resolution images offer more detail, which can be critical in many domains. In addition to that, pattern recognition and other computer vision tasks can be improved with high resolution images. Fig. 3.1 shows the relation between a high resolution image and a low resolution image, where the high resolution image has a larger grid with higher pixel density and the low resolution is obtained by applying an averaging operation on the pixels of the high resolution image [41]. Super resolution may be sought for various reasons and imaging modalities, such as satellite images, security camera footage for facial recognition, number plate identification or also obtaining print quality data [41]. In medical imaging, high resolution images facilitate diagnosis and increased detail is required to delineate volumes, differentiate scar tissue from healthy tissue, amongst the multitude of possible image analysis problems. Super resolution also allows to

save time at acquisition stages by providing equivalent image quality for shorter acquisition times. Super resolution is possible to achieve for single images as well as multiple images. The paradigm of super resolution is that images are intrinsically obtained using a sub-optimal imaging device on a high resolution object. The argument is that every pixel in a low resolution image is a representation of a real scene subjected to decimation and blurring. Using knowledge of the imaging process allows to reverse engineer the blurring and decimation by setting the super resolution problem as an inverse problem.

Those kind of inverse problems are ill-posed, meaning that they have no unique solution and they can only be solved assuming a certain model that can be general or learned from a database. The acquisition is modelled using a combination of a point spread function (PSF), geometrical transformations as well as a noise model to reproduce the decimation, blur and noise inherent to the imaging process. In Fig. 3.1, the PSF is modelled by simple averaging. It has to be noted that in this example any one low resolution pixel has more than one possible match for the values of the four high resolution pixels, leading to the inverse problem being ill-posed as we just mentioned (More on the acquisition model in section 3.2). These super resolution approaches can also called deconvolution, because an assumption is made that the imaging process is similar to a convolution with a PSF and solving the inverse problem to recover the original scene is hence called deconvolution. The methods in this section are mostly based on general models. Super resolution of sets of images, or multi frame super resolution, comes in different forms. It concerns exploitation of images from different modalities, angles or views, colour channels, as well as adjacent frames in video sequences. The overlap obtained from additional data is used to generate a higher resolution version.

3.2 Super resolution and reconstruction methods in MRI

MRI acquisitions should aim to be rapid whilst providing a good signal-to-noise ratio (SNR) and high resolution to allow clinicians to distinguish structures and features in the image. Acquiring images at higher resolution increases imaging time and lowers the SNR. Image quality depends on SNR, and long acquisition times are to be avoided for several reasons. Reducing the time spent in the scanner reduces the possible motion artifacts, makes the process more tolerable for claustrophobic patients and increases the possibility for additional scanner use. Another important factor is the cost effectiveness: scanner time is expensive and needs to be kept to a minimum. Imaging time can be decreased at acquisition stage via smart partial k-space sampling strategies, also called compressed sensing [42] or at post-processing with methods such as super resolution. Sampling strategies in compressed sensing explore different sampling methods of the k-space, where for the same image fewer samples are acquired leading to faster imaging at some cost in image quality. As a matter of fact, cardiac MRI can be accelerated by compressed sensing, where the same imaging planes may be acquired, just by sampling less of the k-space and thus requiring less scanner time. This technique may be used in patients who cannot hold their breath long enough, it does however comes with a loss in image quality [43]. Super resolution in medical imaging and more specifically for MRI is a very active field of research. Most super resolution algorithms use image acquisition models that set up a mapping between the high resolution real scene and the low resolution observation and perform iterative optimisations based on these. A general approach of modelling the acquisition process is to assume the low resolution images are degraded versions of the original high resolution image, related by a down-sampling, low-pass filtering, and noise adding process. A general image acquisition model may be formulated as follows:

$$Y_k = (f_k(X) * h) \downarrow s + v_k; \qquad k = (1, ..., N) \tag{3.1}$$

Where X is the original high resolution image, Y_k is the k-th low resolution image, f_k is the k-th transformation relating to Y_k. The transformation is convolved with h, a low-pass kernel before being down-sampled as represented by $\downarrow s$, N is the number of low resolution observations and v_k is the noise added. h can be seen as the PSF, and depends on the imaging modality. In super resolution algorithms, the PSF is assumed to be known, but this information is difficult to obtain. PSF may be estimated by imaging phantoms where the ground truth is compared to the acquired image. In MRI, the PSF is often approximated to a 3D anisotropic Gaussian [44] and the full width half maximum of the PSF is taken as the slice thickness of the MRI slice. The acquisition model can be generalised as:

$$Y_k = T_k X + v_k; \qquad k = (1, ..., N) \tag{3.2}$$

Where T_k is a transformation incorporating a geometric transformation, low pass filtering, and down-sampling. The acquisition model in Eq. (3.2) can be used to evaluate a high resolution image from several low resolution observations using several approaches. This is at the foundation of the iterative back-projection (IBP) method which was proposed in [45], a seminal publication in super resolution in which an error function related to the current low resolution projections and their differences to the observed low resolution is minimised. It follows the same idea as CT back projection reconstructions, in which the PSF can be estimated from imaging a phantom. However, because it is an ill-posed inverse problem, it does not have a unique solution when using the IBP, which has led to other algorithms relying on prior knowledge about the solution. These other approaches include applications such as maximum likelihood (ML) methods [46] and maximum a posteriori (MAP), which are both also based on solving an inverse problem incorporating an image acquisition model. ML methods aim to produce a high resolution image maximising the probability of low resolution observations and do not rely on prior information about the final output like the MAP estimators do, as

their name indicates. MAP methods use a priori information about the image to be super resolved. This information is added as a constraint to the equation that is being optimised. In several applications [47,48] a smoothing term, or regulariser, is added as a constraint to the optimisation problem one seeks to minimise, and weighted by a Lagrangian multiplier as follows:

$$\tilde{X}^* = \arg\min_{\tilde{X}} \sum_{k=1}^{N} ||T_k\tilde{X} - Y_k||_2^2 + \lambda R\tilde{X} \tag{3.3}$$

Where the first term ensures data fidelity between the current estimate $\tilde{X}$ and low resolution observations, λ is the Lagrangian multiplier, a weighting constant determining the amount of contribution of R, a regularising term. A too high or too low λ may lead to over or under regularisation, a good balance is often evaluated empirically. Regularisers used in MRI MAP reconstructions are often a norm of an operation such as the first or second order derivative [47] or total variation, and promote smoothness in the reconstructed data. Total variation looks at finite differences between pixels, enforcing smoothness by suppressing high differences in adjacent pixels. Total variation is a popular regulariser with edge-preserving and convex characteristics, and has especially been used for de-noising [49]. It has also been applied to super-resolution and for reconstruction of compressed sensing MRI data as a regulariser [50].

There also exist image super resolution techniques not relying on multiple low resolution images, called single image super resolution. While multiple view and single view super resolution both go under the same name, the inherent idea is different. Single image super resolution methods aim to improve resolution post processing without acquisition models used to generate multi-slice images of different orientations creating overlap between slices and checking data consistency against the observed data. Instead, the idea is to reconstruct high resolution images using the low resolution image content. Single image super resolution may be achieved via coupled dictionaries [51] where correspondence between low resolution and high

resolution image patches is sought. A review on single image super resolution of natural images can be found in [52]. Work has been extended to MRI [53] and shown to produce results outperforming interpolation on brain MRI.

The general framework that has been used for super resolution from multiple images in the literature is presented above. We are focusing more on MAP estimation methods here and abstain from introducing methods using projections onto convex sets (POCS), which have very slow convergence speed according to [54]. Early super resolution methods in MRI such as Greenspan *et al.* [55] relied on exploiting additional information from sub pixel shifts parallel to the slice selection direction. Multi-slice MRI acquires a number of slices with anisotropic resolution where the through plane (slice selection direction) resolution is much worse than the in plane resolution, often by large factors. Introducing sub pixel shifts in the slice selection direction allowed to recover information in the through plane resolution and therefore an isotropic reconstruction. Carmi *et al.* [56] offered another study in which super resolution reconstruction of several images of phantoms with different fields of view introducing sub pixel shifts is presented. Later efforts including [57] proposed to introduce sub pixel shifts from a rotational point of view and not just in the slice selection direction. Their results show resolution enhancement on a phantom, ex-vivo and in-vivo clinical brain MRI. An illustration showing pixel shifts in the slice selection direction and rotational sub pixel shifts is shown in Fig. 3.2. While rotation may be arbitrary, pixel shifts resulting from such rotations about an axis may be compared to combining different orthogonal views, in standard MRI they are sagittal, coronal, and transversal. It may be noted that this standard nomenclature does not apply to cardiac MRI, where short axis and long axis including HLA, VLA and LVOT are generally acquired, where all of the long axis are roughly orthogonal to the short axis. Orthogonal views have the advantage of keeping the redundancy within different scans low, a very important feature when scanning times are to be kept at a minimum. Parallel as well as rotated scans have proven to increase through plane resolution, however when using

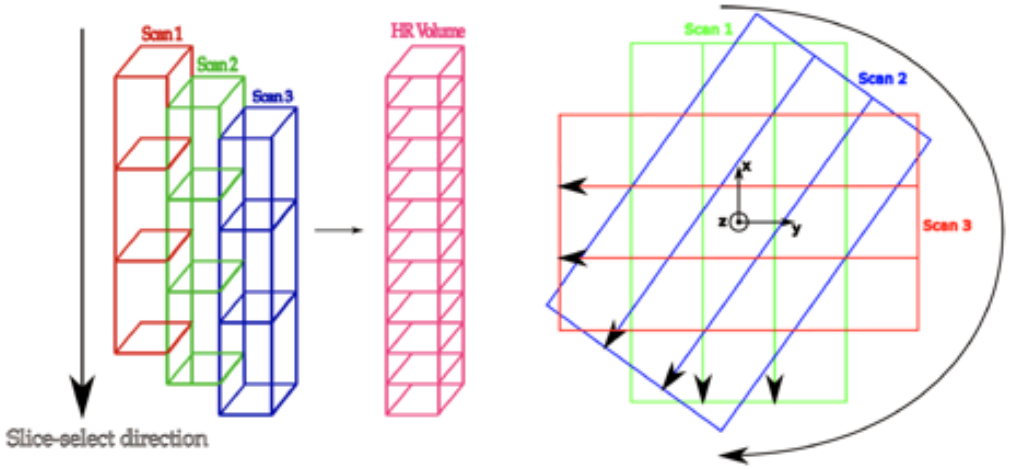


Figure 3.2: Pixel shifts in slice selection direction and resulting from a rotation. Adapted from [54].

the same number of scans, rotated scans outperform parallel scans [57]. The same authors discuss the difference between parallel and multiple view strategies in [58], notably that using multiple views allows a better recovery of spatial frequencies over parallel acquisitions. The reconstructions from rotated scans are much sharper and detailed compared to the ones from parallel scans.

The framework of using regularisation on multi-slice MRI from non-parallel slices has been applied to MRI in [47], which is a study comparing iterative back projection, algebraic reconstruction and two regularised least squared methods. In [47], the authors focus on the reconstruction of phantoms as a preliminary stage before application on a post mortem mouse and ex vivo rat knees allowing static high resolution acquisitions for validation. The quantitative analysis shows that the super resolution algorithms outperform direct high resolution acquisitions in terms of signal to noise ratios and other quality metrics but the performance of the algorithms could not be ranked as they performed variably on the different acquisitions. In Poot *et al.* [48], MRI images of a bird are reconstructed using a MAP estimation where the regularisation term is taken as the square of the laplacian of the current estimate. Results showed combining thirty-six different orientations produced isotropic 3D resolution. They also extended their study to diffusion weighted MRI [59] using the

same optimising function and regulariser. Total variation in addition to an edge enhancing regulariser in the up or down-sampling step, is used as a regulariser in [50] for brain MRI super resolution, however the through plane and in plane resolution difference with $1 \times 1 \times 1.25mm^3$ was not as severe as it is in 2D multi-slice cardiac MRI. One aspect all these studies have in common, is that they outperform standard interpolation techniques, validating super resolution in an MRI context. In addition to that, the regularisation parameter always had to be estimated empirically. Other similar approaches on MRI and diffusion tensor MRI super resolution can be found here [60–62].

An added difficulty in cardiac MRI are the differences in intensity and contrast between short axis and long axis acquisitions occurring because of different imaging protocols and sequence timings. Thus, any reconstruction algorithm aiming to reconstruct images using information from both the short axis and long axis images acquired in standard clinical cardiac multi-slice MRI must be contrast independent. Directional total variation is a recently introduced approach for reconstructing images using a reference image with the same structure but different contrast. In a study by Ehrhardt *et al.* [63] 2D directional total variation is used to combine information from brain MRI with different T1 and T2 weightings. It uses one image and its structural information as a reference for the directional total variation to guide the reconstruction of the other. In Chapter 4, we use a 3D version of directional total variation as a regulariser used for cardiac MRI reconstruction from multiple slices.

3.3 Super resolution in Cardiac MRI

In cardiac MRI, as has been noted before, imaging views are different from conventional MRI in other organs. This is principally due to the main cardiac axis not being aligned with any of the main axes of the body. Short axis stacks of approximately ten slices in addition to fewer orthogonal long axis slices (1-3) are acquired. The short axis image stacks often do not cover the entirety of the heart,

and there are often gaps between slices as can be seen in Fig. 2.6, meaning that there is no overlap between the short axis slices. The only overlap is between the long and short axis slices. Up-sampling along the slice selection direction and super resolution may be seen as interchangeable terms. As shown in [58], multi-slice MRI protocols with rotated, non-parallel views give better results for super resolution compared to the same MAP algorithms fed with data in the form of parallel images. Early occurrences of super resolution reconstruction applied to cardiac MRI include a method based on averaging very large amounts of parallel slices and performing misalignment correction using registration [64]. Another early example was a multi-frame technique extracting high frequency information from adjacent frames in the sequence [65]. Using a wavelet based POCS approach, the authors propose a method that up-samples from size 256×256 to 512×512 and claim the results are qualitatively superior to cubic spline interpolation.

With the MRI super resolution state of the art moving to MAP in the mid to late 2000s, those approaches were also applied to cardiac MRI, with Rahman *et al.* [66] being the first to apply the framework that was initially applied to brain MRI [62]. In [66], the authors produce cardiac MRI super resolution using MAP estimation with the aim to improve segmentations of the left ventricle. In their paper, preliminary evaluation of the super resolution technique on synthetic data is described. This was done to check whether the high degree of anisotropy does not disqualify cardiac MRI data. Indeed, in [62], the low resolution data is $1 \times 1 \times 1.5mm^3$ and much less anisotropic than in cardiac MRI. The results on the synthetic data match the ground truth qualitatively and therefore the authors conclude the validity of the method. The cardiac images they use consisted in two volumes, one in the short axis orientation and one in a long axis orientation orthogonal to the short axis volume, where both volumes have a resolution of $1 \times 1 \times 8mm^3$. Pre-processing of the data was done, including registration, alignment and cubic interpolation to initialise to the image to an isotropic grid. Authors did not perform quantitative evaluation because of the lack of ground truth, however their results compared to

plain averaging show an increase in homogenisation of intensities, sharper contrast as well as enhancement of small structures such as papillary muscles and valves. Improvements are also shown on left ventricular segmentation results. Their method did not include the third long axis view because of long computing times. However, the same authors produced another study [67], in which super resolution is achieved by combining an additional orthogonal view to the previously used long axis and short axis volumes. The authors confirm that a combination of two views (VLA and short axis) outperforms simple averaging between anisotropic slices while interestingly the addition of a third long axis view does not seem to further benefit the reconstructions. They argue this is because of the high degree of sensitivity of MAP to the alignment of the images. Even with motion correction, some degree of imperfection is likely to persist, and it will only be amplified when using three instead of two volumes. Both their studies use MAP estimation based super resolution and produce promising results, underlining the validity of using both short axis and long axis information from cardiac MRI. Another MAP approach is explored in [68]. A Wiener filter is used in addition to the MAP estimator, to remove additive noise resulting from the multiple low resolution images being combined. Results on synthetic data show qualitative improvements, and small quantitative improvements over using simple interpolation. While the authors claim to have performed pixel intensity correction on the clinical data, the results clearly display intensity differences occurring where the short axis slice informations were used in the long axis view, and where the long axis information was used in the short axis view. In [69], super resolution of Gadolinium enhanced images is done with the aim to better quantify the sizes of scar tissues post myocardial infarct. The data used is similar to the volumes from [66, 67], with the difference that an overlap of half the slice width is introduced. Super resolution is achieved by applying the methods used in [48] which are extended to include alignment parameters of the image obtained from pre-processing registration between the images. The authors used a 3D sinc point spread function to model acquisition and initialised the image to be overly smooth before iteratively optimising it, and visually decided upon the

final reconstruction of which they carried the parameters forward to other datasets. Results show adding the long axis images outperforms using the short axis stack for reconstruction alone, validation is performed on electroanatomical mapping that was performed after intervention. In [70] a reconstruction of cardiac cine MRI sequences is proposed, using a method based on iterative reconstruction and registration. The k-space is synthetically prospectively under-sampled for all but the first of the frames of the sequence, and that first frame is used as reference image for each of the other frames independently to give results with good visual quality. This study, however, did not treat the images as being specifically cardiac MRI, it focused more on the k-space under-sampling of an MRI acquisition.

All but one of the previously mentioned methods for cardiac MRI work on static images. In work by Odille *et al.* [71] cine MRI sequences of 32 frames covering the left ventricle in short axis, HLA and VLA orientations are used to reconstruct a 3D volume. Each orientation is covered by stacks of 20 images, similar to [66, 67, 69]. The authors of [71] use temporal information for motion compensation. The imaging protocols used in those studies do not follow clinical practice, which makes sense as clinical cardiac MRI is not acquired to maximise overlap between slices having super resolution reconstruction in mind. It is noted that the reconstruction times are quite long, being reported as over 3 hours for the motion component alone. The authors evaluated the results visually and by drawing intensity profiles, and the profiles show improvements around structures like the myocardial wall against the observed low resolution data.

Simultaneous to the establishment of MAP as a temporary state of the art for super resolution reconstruction in MRI, other methods relying on prior data (also called example based) have emerged and been applied successfully to cardiac MRI. Example based techniques include dictionary learning and atlases. The aim of atlas based methods is to recover high resolution patches from low resolution ones by matching corresponding patches between an the image and an atlas. Methods based

on the use of atlases, are computationally expensive because the patches closest to the patches of interest have to be found in a high resolution database. In [72], a novel super-resolution approach using the multi-atlas PatchMatch algorithm is presented in which a cardiac 3D volume is generated using only a short axis stack. It is based on learning patches from a database and requires vast training data, like in a study from the same group [73] where an atlas was built for shape, motion and wall thickness analysis using over a thousand 3D SSFP images. The study is based on 189 3D images as atlases and 11 low resolution images for evaluation, they are synthetically generated by down-sampling 3D images. Results show considerable improvements when compared to linear and cubic interpolations. Another recent effort on reconstruction of cardiac MRI using atlases has been proposed in a framework using 3D patch-based self similarity where stacks of 20 anisotropic images acquired in short axis, HLA and VLA views are interpolated and combined [74], in addition to demonstration on phantom and brain data. This study uses a protocol maximising coverage of the heart and overlap between slices similar to the one used in the study previously mentioned [71]. The drawback is that this protocol is not followed in clinical practice and much more time consuming since the total number of acquisitions is approximately increased threefold. In a recent paper by the same group [75], authors build on their recent work on cardiac MRI super resolution and motion correction to compare manual segmentation strategies. Their results find that the segmentation of standard short axis stacks is not the optimal way to obtain the ventricular volumes, and that an overall lower number of acquisitions of sparse long axis would give better results.

Dictionary learning approaches to super resolution aim to exploit redundancy in the data, by matching image patches extracted from a low resolution image to high resolution image patches in a high resolution dictionary to upsample a low resolution image. A signal, or image, is said to admit a sparse approximation using a dictionary if there exists a linear combination of elements (also called atoms) from the dictionary, allowing to reconstruct an adequate representation of the initial

signal. A sparse vector is a vector with very few non-zero elements containing the coefficients of a sparse representation, which returns an approximation of the signal when combined with the dictionary used for decomposition. A dictionary can be represented by a $N \times K$ matrix serving as the database of relatively small signal sequences, for example image patches. K is the size of the dictionary determined by the number of atoms, and N is the length or dimensionality of each atom. Off-the-shelf dictionaries, such as ones made up of wavelets, are available. However, they are designed for specific kinds of images with smooth boundaries, and may not be able to handle new arbitrary signals and have limited signal sparsifying abilities, therefore dictionaries are often learnt for specific applications using specific data [76,77]. An example of single image super resolution reconstruction in cardiac MRI is in a recent study by Bhatia *et al.* [78]. The authors co-train low resolution and high resolution 2D dictionaries on synthetic data. Even though short axis views are combined to obtain a low resolution long axis image, super resolution is done on a low resolution long axis view only so it can be counted as single image super resolution. In this study, an approach of pattern matching between high frequency information, such as edges or textures, and low resolution image patches from coupled dictionaries to produce a higher resolution reconstruction of a long axis slice using several short axis slice stacks is performed. The final image is obtained by superimposing an image containing high frequency informations resulting from the dictionary recomposition onto the interpolation of the low resolution images. Results on synthetically generated data achieve a qualitatively impressive result, however the results on clinical data highlight the difficulties coming with real clinical short axis data. 3D dictionaries have successfully been implemented into up-sampling images, for example to accelerate Fourier-Transform Infrared data acquisition times which outperformed standard interpolation [79] or in brain MRI, outperforming conventional reconstruction methods [80]. Dictionary learning has also been applied on the other end of the imaging pipeline to speed up cardiac MRI via compressed sensing in [81], where a dynamic 2D+t dictionary is learnt and used to recover missing k-space data.

3.4 Metrics used for quality quantification super resolution

The two most commonly used metrics for quantitative assessment within image processing, more specifically super resolution and image reconstruction, are the peak signal-to-noise-ratio (PSNR), based on the mean square error (MSE), and the structural similarity (SSIM) [82], which has been designed to improve on the previous metrics. The MSE, and the PSNR which is based on the MSE, are metrics that treat absolute errors, where pixel-to-pixel differences are used [83]. Both the MSE and PSNR are widely used in image quality assessment because of their simplicity and clear physical meaning. However, they are often criticised for not matching visual quality. The SSIM was designed with the aim to correlate the quality of perception of the human visual system by combining luminance, contrast, and structure. Both metrics are used throughout the reconstruction and super resolution literature, especially in literature in which qualitative evaluation is taken to its limit and improvements are best described by metrics. The MSE between a reference image A and a test image taken for comparison B of the same size $m \times n$ is defined as:

$$MSE = \frac{1}{m\,n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [A(i,j) - B(i,j)]^2 \tag{3.4}$$

PSNR builds on MSE using the maximum pixel intensity (MPI) of an array, PSNR is defined as:

$$PSNR = 10 \cdot \log_{10}(\frac{MPI^2}{MSE}) \tag{3.5}$$

Low error between the reference and test image would result in low MSE, and as the MSE approaches zero, following inverse proportionality the PSNR value approaches infinity. A higher PSNR value therefore correlates with higher quality of the test image, whereas a low PSNR would indicate that the differences between the two images are high.

The SSIM aims to improve on MSE and PSNR, as different distortions of the same image can have the same numerical MSE (as shown in Fig. 2 of [82]) but drastically different appearance. SSIM is evaluated on a local scale and not on a global scale, pixel by pixel sliding windows of 8×8 pixels are evaluated. The SSIM is obtained by multiplying formulas for luminance, contrast and structure to yield:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.6}$$

where μ_x and μ_y are the averages of windows x and y in the reference and test images, σ_x^2 and σ_y^2 their variances and σ_{xy} their covariance. C_1 and C_2 are constants related to the dynamic range of the pixel values. SSIM values range from -1 to 1, where a value of 1 is only achieved for two identical images. The quantitative evaluations of the contributions of this thesis are done mainly using PSNR and SSIM.

Reviewing the state of the art literature on super resolution reconstruction approaches in cardiac MRI, many similarities between techniques become obvious. None of the regularised least squares studies dealing with cardiac MRI reconstruction have used real anisotropic multi-slice 2D cardiac MRI data as is typically acquired in clinical settings [35, 37]. Recent deep learning work by Oktay *et al.* [84] has used a clinical 2D dataset. However, this is understandable as most regularised least squares super resolution reconstruction techniques are based on making use of overlap between intersecting slices, and standardised protocols in clinical cardiac MRI do not favour such overlap. Indeed, the data used in studies relating to super resolution reconstruction in the literature is generally data that has been specially acquired for the experiments with the experiment design in mind. Often, stacks of short axis, HLA and VLA are obtained and merged using averaging or MAP estimations to produce enhanced volumes. Also, the methods not relying on the use of numerous additional long axis views simply use the short axis stack, or rather synthetically generated short axis stacks from 3D images [72] or from long axis

slices [78], so the combination of short axis and a handful of long axis is something yet to be explored. In addition to that, only few approaches take advantage of leveraging consecutive cine frames, [65] for simple up-sampling of images and [71] for motion correction of three isotropic volumes or rather image stacks. Lastly, there is no mention of intensity correction in any of the literature we have seen thus far, even though intensity mismatch between short axis and long axis and even within short axis is an issue in 2D multi-slice cardiac MRI.

The project subject of this thesis had an interesting timing, with the shift of the state of the art. From early methods such as IBP, settling to sets of MAP techniques and atlas or dictionary based techniques also showing great promise. Starting with those example based techniques, the state of the art has drastically shifted in the last 2-3 years to the application of deep learning techniques. At the start of the project, MAP techniques were the state of the art, however, with the advent of deep learning and its incredibly rapid expansion, the research undertaken for this project has taken a drastic turn and we present more background and the updated state of the art which has arisen from deep learning in chapter 5 in order to introduce the 6th and 7th chapters of the thesis dealing with our efforts with deep learning approaches.

This chapter gave the reader an overview of super resolution reconstruction techniques of the last few decades with a focus on MRI and cardiac MRI reconstruction studies. We learnt that most techniques rely on good coverage of data, sufficient overlap between slices and alignment of data. At least two of those traits are not descriptive of cardiac MRI data following standard protocols, an issue encountered by us as well. The next chapter describes our efforts on reconstruction of cardiac MRI using a technique based on the state of art described so far.

4

Reconstruction of Three-Dimensional Cardiac MRI Using Directional Total Variation

The previous chapter formed a literature review which included a substantial amount of studies and examples describing traditional super resolution reconstruction approaches, mostly based on regularised least squares. In the following chapter, we introduce a new such method based on directional total variation and also propose an alternative slice planning protocol. We presented part of the work in this chapter as an oral presentation at the Reconstruction and Analysis of Moving Body Organs (RAMBO) workshop of the MICCAI 2017 conference in Quebec City, Canada, and it was published in the proceedings [9].

4.1 Preamble

Initial approach using dictionary learning

The first year of the time dedicated to the research project described in this thesis was focused on the use of dictionary learning towards reconstruction of cardiac MRI. To get familiar with dictionary learning, a first approach was to perform simple up-sampling using coupled dictionaries similar to the previously mentioned [78]. Instead of processing 2D data with coupled 2D dictionaries, the aim was to process 3D data with coupled 3D dictionaries. For training and testing, we used high resolution 3D data from pre-clinical studies we obtained from collaborators. An example of two pairs of atoms from jointly learnt low and high resolution dictionaries can be seen in Fig. 4.1, where the left hand side shows low resolution $8 \times 8 \times 4$ atoms and the right hand side $8 \times 8 \times 8$ higher resolution matching atoms. Both atoms display edge shapes, and it can be seen that the boundaries are more detailed in the high resolution atoms. The low resolution training data was generated by simple down-sampling of the high resolution data by a factor of two. The low resolution and high resolution dictionaries were jointly trained. We performed the dictionary training and the sparse decomposition tasks using the SPAMS (SPArse Modeling Software) toolbox developed by Julien Mairal. The toolbox contains a multitude of functions ranging from dictionary learning algorithms, sparse decomposition solvers to image manipulation functions and more [85]. We synthetically down-sampled the high resolution images by a factor of two in the slice selection direction in order to have low resolution data matching the high resolution data. We prepared the images for training by dividing the high and low resolution 3D images into small 3D blocks before reshaping each block into 1D vectors, as the dictionary learning is done on 1D vectors. After concatenating the high and low resolution arrays, the dictionaries were learnt. This way, the low and high resolution dictionaries have a one-to-one correspondence between themselves and have the same sparse representation for corresponding patches.

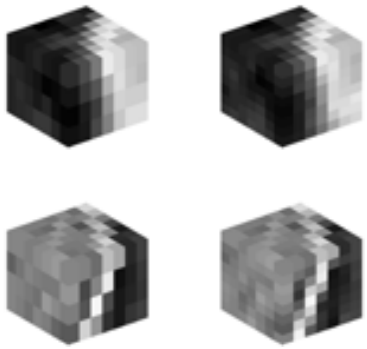


Figure 4.1: Two atoms of coupled low and high resolution 3D dictionaries.

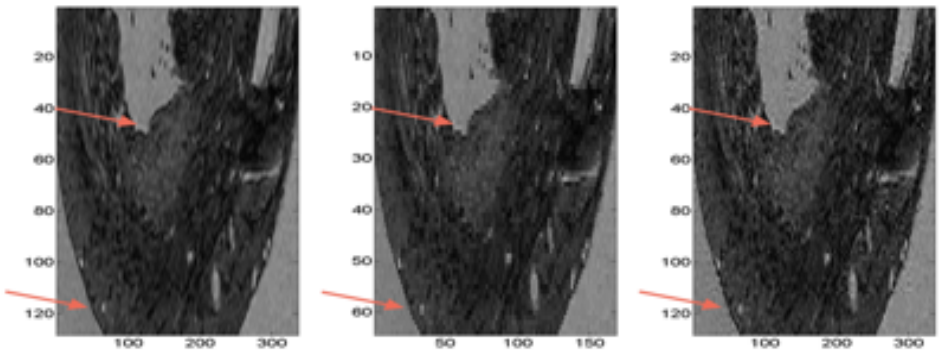


Figure 4.2: Up-sampling of 3D data using coupled dictionaries. Left: ground truth, middle: low resolution, right: result using dictionaries. The red arrows point edge locations where differences are most noticeable.

A result of up-sampling of 3D data by a factor of two can be seen in Fig. 4.2. Here we show a 2D view through the volume. The arrows point to locations of features where differences are most noticeable, showing smoother boundaries in the reconstructed image and are more ragged features in the low resolution image. We produced the high resolution reconstruction by taking a low resolution image, performing sparse decomposition with the low resolution dictionary and combining the resulting sparse vector with the matching high resolution dictionary. We developed code in MatLab used to transform large 3D volumes into small voxels, in order to learn dictionaries and perform sparse decomposition of the signals.

The aim of the initial project layout was to move on to an iterative scheme following a similar approach to one presented in [86], where prospectively under-sampled MRI is accelerated using dictionary learning in an iterative scheme, with some elements performed in the Fourier domain. Their problem formulation includes a term related to sparsity, which is meant to enhance high frequency contributions in the reconstruction by exploiting redundancies in the k-space. For our case, some important limiting factors were the time required for each matrix decomposition, intensity differences between long and short axis acquisitions in clinical data, the lack of high resolution ground truth 3D data, and the fact that we aimed to treat non-parallel data. The time required to learn dictionaries also exceeded several days, on the limited amount of data that we had access to.

We decided to change the focus of our research away from dictionary learning, however the experience with medical image data and the 2D and 3D data handling developed for the project with dictionary learning in mind were very useful and crucial for the rest of the thesis work. After continuous assessment of the literature which showed the state of the art dramatically shifting towards deep learning approaches, and seeing how promising our initial attempts at using these new methods were, the focus of the thesis was focused towards deep learning efforts. One of the few parameters that we theoretically could experiment on with dictionary learning was changing the size of dictionary elements, however when the atoms exceeded $8 \times 8 \times 8$ training became computationally too expensive (more than one week). With deep learning, the options for experimenting with many different architectures and filter sizes became very appealing. Before completely shifting the focus, however, the work started on dictionary learning was modified to become the algorithm presented in this chapter. This gives the thesis a chance to explore a more traditional approach based on regularised least squares and compare it to current deep learning approaches more in line with the research undertaken in the field at this point in time.

4.2 Introduction

We have seen in chapter 3 that super resolution reconstruction post-processing methods in MRI have been the topic of considerable amount of literature over the last two decades, leading to the set of methods before the advent of deep learning, which we have labeled as *traditional*. The majority of these involve an assumption made on an image acquisition model and to reverse engineer the process using iterative regularised schemes. Studies in MRI super resolution reconstruction take advantage of overlap in the data between different orientations of multi-slice anisotropic data, with the early super resolution reconstruction methods exploiting sub-pixel shifts from parallel offsets introduced in the through plane direction [55] where the resolution can often be several factors worse than the in-plane resolution. With super resolution reconstructions in mind, improvements have been made to sampling strategies, and information gains from sub pixel shifts were introduced from rotations of acquisition planes, a process which was shown to outperform parallel shifts when using the same number of images [48, 58]. In cardiac MRI, standardised protocols for multi-slice imaging in clinical practice are in place [35, 37] and images are acquired in parallel short axis stacks with gaps between them and fewer long axis slices orthogonal to the short axis stack. This means, however, that overlap is highly restricted, only occurring at intersections between the short and the long axis acquisitions, which is not ideal for the super resolution approaches described in relevant MRI literature. Indeed, efforts made to produce cardiac MRI super resolution reconstructions do not make use of standard clinical multi-slice 2D cardiac MRI. Several studies describe protocols that differ from the accepted clinical standard, achieving extensive coverage by means of changing acquisitions of long axis slices into long axis stack acquisitions [66, 67, 69, 71]. The literature fails to mention that these multi-stack protocols are significantly different to standardised ones, or the need to change clinical protocols to comply with these super resolution reconstruction paradigms. Sample image acquisition protocols used in the super resolution literature are shown in Fig. 4.3.

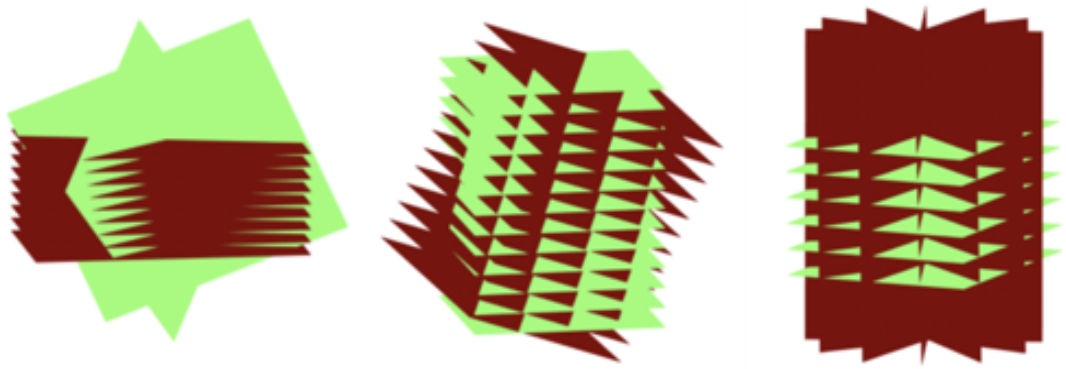


Figure 4.3: Various multi-slice approaches for cardiac MRI coverage. Left: short axis stack in green with a few long axis slices in red (clinical standard), Middle: Stack of short axis in green combined with stack in another orientation in red (as used by e.g. [66,67,69,71]), Right: stack of short axis in green and several long axis slices in red as proposed later in this chapter.

The left hand side of Fig. 4.3 shows how standardised multi-slice 2D cardiac MRI protocols cover the heart, where a stack of parallel images cover the short axis and only a few orthogonal long axis images are acquired. The UK Biobank protocol [35] recommends three long axis slices, and the number of short axis slices is dependent on the size of the heart but is usually around ten slices. The middle image in Fig. 4.3 of shows a multi-stack acquisition, where another stack is acquired in addition to the commonly acquired short axis stack (as used by e.g. [66,67,69,71]). Here, the total number of slices is doubled or tripled compared to standardised protocols. The right hand side shows how a protocol balancing the number of short and long axis slices could look like. Standard long axis acquisitions that cut approximately through the middle of the heart are of clinical validity, as they include features of the heart such as the base and apex which are not well covered, or may be missed, by the short axis stack. In addition to that they can help as an additional reference for slice misalignment correction, which can be similar to

the banana problem encountered in histology[1]. However, introducing additional images parallel to the original long axis slices may not necessarily add much to clinical interest, such as the outer slices from an additional stack as shown in the middle of Fig. 4.3 which may only capture a minimal part of the myocardium. For this reason, we show a proposed slice protocol with additional rotated instead of parallel long axis slices and balancing the absolute number of short and long axis slices, instead of acquiring stacks of long axis images. We have seen how others have addressed the task of reconstruction of 2D multi-slice cardiac MRI in section 3.3 and present our own approach to the problem in the following chapter which proposes a different imaging protocol than the literature and offers a way to treat multi-slice data in a contrast independent manner in 3D.

4.3 Data

In this first part of the work undertaken for the thesis, two different kinds of data have been used. We present the pre-clinical low resolution 2D images and high resolution volumes used for for prototyping and validation of the algorithm, as well as the clinical 3D volumes used to generate multi-slice 2D acquisitions used to show potential applications and shortcomings of the method.

4.3.1 Pre-clinical data (ex vivo)

To prototype and validate the method, we use pre-clinical data from excised rat hearts accompanied by corresponding high resolution ground truth volumes that were acquired specifically for our project by collaborators from the team of Prof Jurgen Schneider at the BHF Magnetic Resonance Unit in Oxford including Dr Irvin Teh and Dr Darryl McClymont. Both high resolution 3D volumes and low resolution 2D multi-slice MRI were acquired, the latter aiming to reproduce standard clinical protocols with highly anisotropic images and gaps between short axis slices. The

[1]The banana problem is the phenomenon that would occur if a banana is sliced and the sliced are then aligned to each other using in-plane transformations. The original curvature of the banana would be lost to produce a shape resembling a cylinder. The use of an additional reference in the long axis would prevent that.

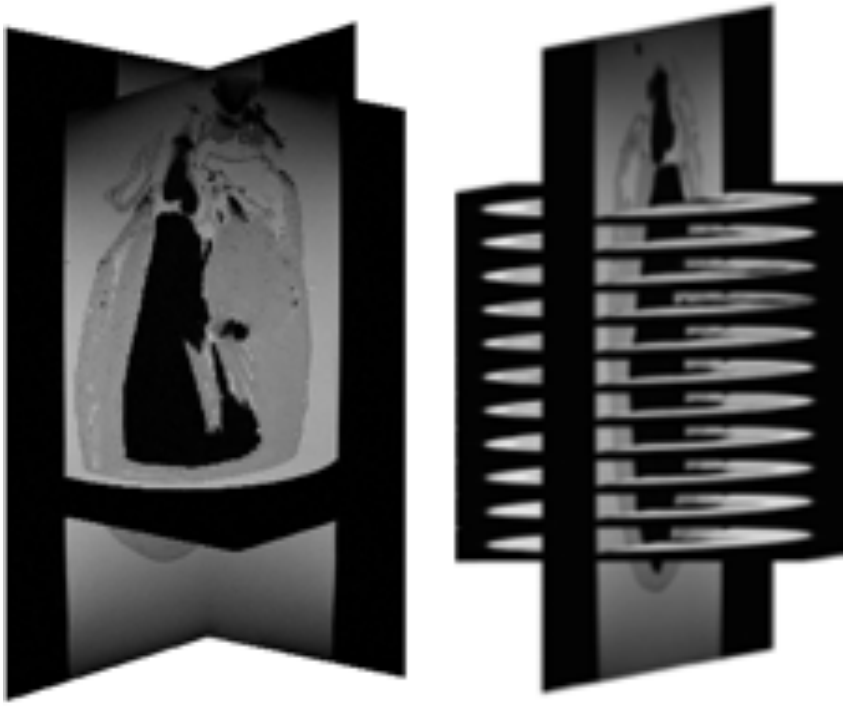


Figure 4.4: Left: Cut through a high resolution pre-clinical data volume, right: 2D low resolution short axis slice stack and one long axis slice of the same excised rat heart.

additional long axis views spread out like a fan, and are not additional parallel offsets of the original long axis views as has been done in other instances in the literature as mentioned above.

Experimental investigations conformed to the UK Home Office guidance on the Operations of Animals (Scientific Procedures) Act 1986 and were approved by the University of Oxford ethical review board. Ten hearts were excised from female Sprague-Dawley rats during terminal anaesthesia, fixed then embedded in 1% agarose gel, and imaged on a 9.4 T pre-clinical MRI scanner (Agilent, CA, USA). A 2D multi-slice gradient echo sequence was used: FOV = 25.6×25.6 mm^2, acquisition matrix = 128×128, slice thickness = 1 mm, slice gap =0.25 mm, TR = 200 ms, TE =2 ms, flip angle = 20°. 11 short axis and 11 long axis slices were acquired in 7.2 min (this can be seen in Fig. 4.5 as well as in the right hand side of Fig. 4.4). For each of the hearts, a single 3D gradient echo image was acquired in the same

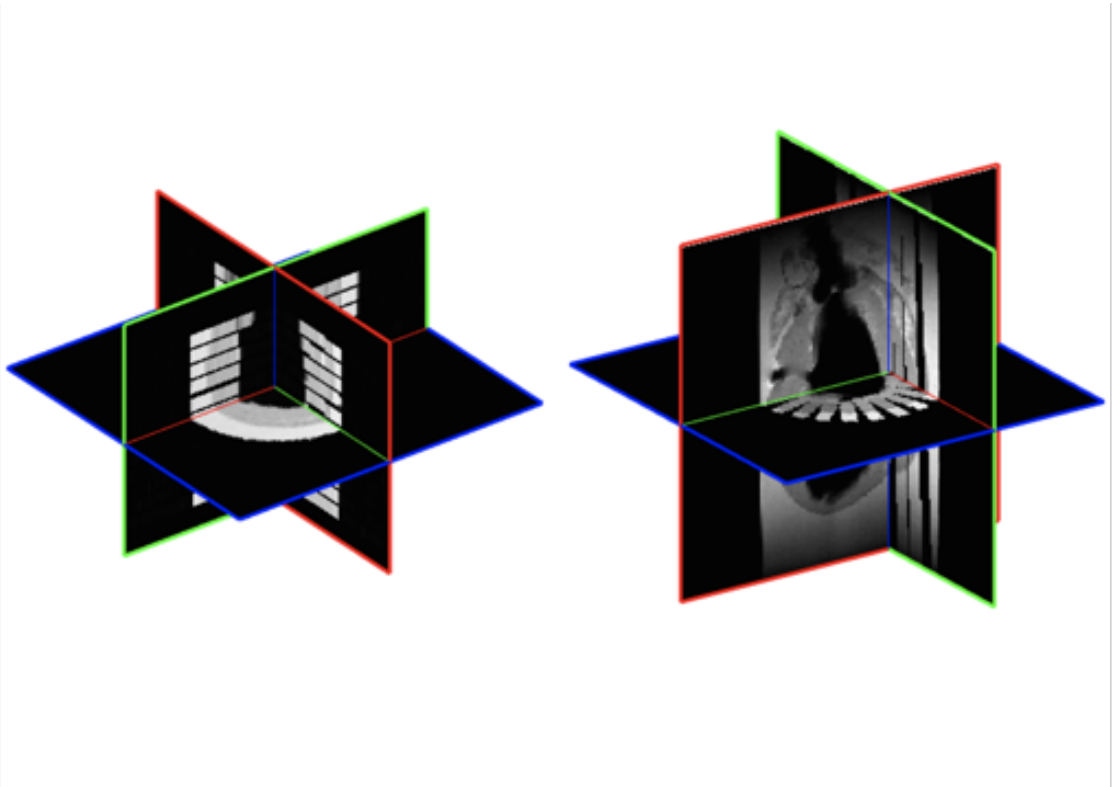


Figure 4.5: Left: cut through 3D view of 11 2D synthetically generated short axis slices resampled into 3D space. The blue plane is aligned with the short axis view while the red and green planes are aligned with long axis views. Right: cut through 3D view of 11 synthetically generated 2D long axis slices resampled into 3D space. The long axis slices have noticeably different contrast than the short axis slices, the myocardium is darker and has lower average intensity values. Adapted from [9].

session: FOV = $25.6 \times 25.6 \times 25.6$ mm^3, acquisition matrix = $384 \times 384 \times 384$, TR = 200 ms, TE = 4 ms, flip angle = 60°, scan time = 8.2 hours (this can be seen on the left hand side of Fig. 4.4). In addition to the real acquisitions, low resolution 2D slices $\mathbf{X}$: FOV = 25.6×25.6 mm^2, acquisition matrix = 128×128, in-plane resolution = 0.2 mm, slice thickness = 1 mm were synthetically generated from the 3D image $\mathbf{Y}$ using a sampling function $\mathbf{T}$, such that $\mathbf{X}_i = \mathbf{T}_i\mathbf{Y}$. The sampling function to generate the low resolution data differs to the ones generally used in the literature by working in k-space and is presented in section 4.4.2. The long axis images were synthesised after applying a histogram shift to the ground truth 3D volume, to ensure the low resolution short and long axis images have different contrast [9]. It may be worth noting that what can be seen in the pre-clinical images is a test tube filled with gel in which the excised hearts are suspended.

4.3.2 Clinical data (in vivo)

The data from the HVSMR 2016 MICCAI workshop on whole-heart and heart vessel segmentation from 3D cardiovascular MRI in congenital heart disease is used to test the algorithm on real human data with ground truth volumes to evaluate the performance of the method. The data consists of 10 volumes acquired using the 3D SFFP sequence. The subjects have a variety of congenital heart defects and were imaged on a 1.5 T scanner (Phillips Achieva) without the use of contrast agents, TR = 3.4 ms, TE = 1.7 ms, flip angle = 60°. The average acquisition matrix size was $390 \times 390 \times 165$. The in-plane resolution for this data is $0.9 \times 0.9 \times$ mm^2 and the slice thickness of 0.85mm. While near-isotropic resolution is achieved in those images, they are noisy and display lower SNR than 2D multi-slice data normally acquired in practice. This data was chosen because of the availability of ground truth for 3D volumes, the fact that is was freely available, and because we could use an image acquisition model to generate low resolution 2D images in any view we wanted.

4.4 Methods

In this chapter, we address the problem of reconstructing 3D images from a combination of a short axis 2D slice stack and long axis slices, in a contrast-independent manner using the directional total variation regulariser. This allows a reconstructed image with the contrast of the short axis images but with the additional structural information contained in the long axis images. In super resolution MRI reconstruction literature, the imaging process is generally modelled as follows: A ground truth object $\mathbf{Y}$ is imaged by a process resulting in one or several observed images, $\mathbf{X_i}$. This is modelled by applying a transformation $\mathbf{T}$ and additive gaussian noise n. $\mathbf{X_i} = \mathbf{T_i Y} + n$. The operator $\mathbf{T_i}$ is defined as a combination of geometric transformations, convolution with a point-spread function which is often a Gaussian kernel, and a downsampling operation [54]. The geometric transformation consists of a rotation of the 3D image, where the array is rotated such that the image plane to be acquired is in the XY orientation. Having defined



Figure 4.6: 6 short axis and 6 long axis slices synthetically generated from HVSMR dataset, patient 17. The artificial intensity difference between short and long axis images is clearly noticeable. The result for this particular reconstruction of multi-slice data is shown in Fig. 4.11.

an acquisition model, the image reconstruction process can be posed as an inverse problem. In addition to the data consistency term, different regularisers R have been applied to guide reconstructions [48, 60]. Such regularisers are based on prior knowledge or assumption of the image and control features of the reconstructions such as the magnitude of edges and degree of smoothness, and allow reconstruction problems to be posed as follows where the reconstructed image is found by solving:

$$\mathbf{Y} = \arg\min_{Y} \sum_{i=1}^{N} ||\mathbf{T}_i\mathbf{Y} - \mathbf{x}_i||^2 + \lambda R \tag{4.1}$$

Where $\mathbf{Y}$ is the 3D volume, $\mathbf{X_i}$ the 2D observations, N their number, and λ determines the weight of the regularisation term, R. λ is often evaluated and

fine-tuned empirically. For a function f, the expression $\arg\min_x f$ corresponds to the input value x which gives the minimum value for f, $\min f$.

4.4.1 Super-resolution algorithm

The approaches described before assume the compatibility of data obtained in different views, which may not always hold especially with misaligned and intensity mismatched slices at locations of overlap. This can be seen in Fig. 2.7 which may not be affected by excessive misalignment, however it displays strong intensity differences between long and short axis views, so it does not make sense to initialise such data by treating the short and long axis slices into the same volume. To address this issue, we treat short and long axis data separately before their combination. Fig. 4.5 shows the first step in the initialisation of two volumes, one using the short axis stack and one using the long axis acquisitions with noticeable contrast differences between the two. We formulate the problem by simultaneously solving the following:

$$\mathbf{Y}^*_{\mathrm{SA}} = \arg\min_{\mathbf{Y}_{\mathrm{SA}}} \sum_{i=1}^{n_{\mathrm{SA}}} ||\mathbf{T}_i\mathbf{Y}_{\mathrm{SA}} - \mathbf{X}_i||_2^2 + \lambda J(\mathbf{Y}_{\mathrm{SA}}, \mathbf{Y}_{\mathrm{LA}}) \tag{4.2}$$

$$\mathbf{Y}^*_{\mathrm{LA}} = \arg\min_{\mathbf{Y}_{\mathrm{LA}}} \sum_{j=1}^{n_{\mathrm{LA}}} ||\mathbf{T}_j\mathbf{Y}_{\mathrm{LA}} - \mathbf{X}_j||_2^2 + \lambda J(\mathbf{Y}_{\mathrm{LA}}, \mathbf{Y}_{\mathrm{SA}}) \tag{4.3}$$

In both the equations 4.2 and 4.3, the first term is related to data accuracy, its aim is to ensure that the current estimate does not deviate too much from the 2D low resolution observation $\mathbf{X}_i$ which are the ground truth measurements. The second term sets a constraint using the directional total variation of the image. It pushes the first term in J towards being smooth whilst using the structural information of the second term in J as a reference. λ is a weight adjusting the contribution of the directional total variation term.

Total variation

Total variation is a popular regularisation functional with desirable properties: it is convex, edge preserving and promotes smoothness and is often used for denoising. It is the sum of finite pixel-to-pixel differences, its formulation for data q of dimension N is as follows:

$$TV(q) = \sum_{n=1}^{N} |\nabla q_n| \tag{4.4}$$

Directional total variation

Total variation in its standard formulation does not allow additional knowledge to be used as a prior. Directional total variation is a recently introduced approach which does allow incorporation of a priori knowledge about the solution. It can be applied to image reconstruction problems by using a reference image with the same structure but different contrast. In the study by Ehrhardt *et al.* [63] a 2D version of directional total variation is used to combine information from brain MRI with different T1 and T2 weightings. Their method uses one image and its structural information as a reference to guide the reconstruction of the other, by exploiting the expected redundancy in the multi-contrast data. Indeed, a T1 weighted image of a certain scene will contain most of the information related to edges and their directions as a T2 weighted image of the same scene would contain as well. The sign of the direction of an edge does not matter, as in one image a jump to a higher pixel value may correspond to a jump to a lower pixel value in another image with different contrast. In the conclusions of the study, the authors note that using the a priori information obtained from direction of edges outperforms conventional total variation as well as weighted total variation, which makes use of the locations of edges alone. We aim to use directional total variation using one of the views to control the anisotropic reconstruction of the other view. In the equations below, A and B correspond to the short and long axis data.

The 3D directional total variation constraint J, as used in equations 4.2 and 4.3, applied to image $\mathbf{A}$ with reference image $\mathbf{B}$ of dimension n is as follows:

$$J(\mathbf{A}, \mathbf{B}) = \sum_{n=1}^{3} |D_n \nabla \mathbf{A}_n| \tag{4.5}$$

where matrix field $D_n \in \mathbb{M}^3 = \mathbb{1} - \xi_n \xi_n^*$ and $\xi_n := \frac{\nabla \mathbf{B}_n}{|\nabla \mathbf{B}_n|_\eta}$, $\mathbb{1}$ is a unit vector field and $*$ denotes a complex conjugate transpose. In which $|\nabla \mathbf{B}_n|_\eta = \sqrt{|\nabla \mathbf{B}_n|^2 + \eta^2}$ for $\eta > 0$. This gives $0 \leq \xi_n < 1$ where the minimum is obtained when the gradient it zero, meaning $\mathbf{B}$ is constant and $|\nabla \mathbf{B}| = 0$. The maximum is approached as $|\nabla \mathbf{B}| = \infty$. The tuning parameter η relates to the size of the edges in reference image $\mathbf{B}$ and controls what value of an edge is considered more or less significant. In data which is normalised and has intensities between 0 and 1, a 5% change can be seen as significant, so we picked $\eta = 0.05$ as a starting point for fine-tuning.

Equations 4.2 and 4.3 were simultaneously solved using nonlinear conjugate gradient optimisation [87], in which one step towards the minimum in both equations 4.2 and 4.3 was taken during each iteration. The image $\mathbf{Y}_{\mathrm{SA}}$ was initialised by putting the low resolution short axis slices into their respective orientations in a 3D matrix with isotropic spacing matching the low resolution in-plane resolution, and performing trilinear interpolation to have a better, but still simple, starting point than nearest neighbour interpolation. $\mathbf{Y}_{\mathrm{LA}}$ is initialised in similar fashion, by putting the low resolution long axis slices into a 3D matrix and filling the gaps using trilinear interpolation. We averaged voxel values where the long axis slices overlapped. The nonlinear conjugate gradient method alternates between iterations as to which image to reconstruct and which one to use as a reference for the directional total variation functional. The λ parameter was empirically set to 0.5 and the η parameter was kept at 0.05. The data consistency module, the first term in equations 4.2 and 4.3, checks observed short axis low resolution against the short axis reconstruction $\mathbf{Y}_{\mathrm{SA}}$, and long axis low resolution against the long axis reconstruction $\mathbf{Y}_{\mathrm{LA}}$. The process is repeated until convergence. Results are shown on the short axis reconstruction

as the aim is to increase their through-plane resolution guided by additional shape information contained in long axis slices.

The 3D reconstructions were validated against the high resolution ground truth volumes that were acquired for that purpose. We used PSNR, SSIM for image quality quantification, and further validated the results on segmentations using percentages of voxel misclassification. The ground truth was taken as the volumes with the contrast from which the short axis images were generated, as such we compared the reconstructions from short axis initialisations to the ground truth. We decided on this as it would not make sense to compare two different contrasts, and the short axis stack is used in standard clinical practice to cover the entire heart volume. Segmentations were obtained by binarising the images via simple thresholding. The contrast between the myocardium and the left ventricle in the pre-clinical data is sufficient that the segmentation result is insensitive to minor changes in the threshold value. The experiments were run using 11 short axis and 11 long axis slices achieving an extensive coverage of most of the space, and then with a total number of 12 slices with different combinations of numbers of short axis and long axis acquisitions in order not to use more slices than commonly acquired in practice when following standardised protocols, which in theory would not require any additional scanner time. We chose the combinations to be 9 / 3, 6 / 6, and 3 / 9 short / long axis slices respectively. For the clinical data we kept the slice protocol of balanced slice numbers which gave the best results for the pre-clinical data and evaluated the PSNR and SSIM only, as segmentation is not achievable by simple thresholding since those images contain more complex features.

4.4.2 Image acquisition models

Synthetic acquisitions in Fourier space

For the generation of synthetic low resolution slices from the volume, we designed an acquisition model in the Fourier domain. The Fourier domain represents spatial frequencies, where the low frequencies related to global image appearance are near

the centre of the Fourier space, while the high frequencies provide edge and detail information and are on the periphery of the Fourier space. Low resolution MRI scans leave out some of the high frequency space. Leaving out the periphery can be done by applying a square or rectangular mask to the sampling space. The pre-clinical data, as presented above, was acquired specifically for our project. From discussions with collaborators in charge of MRI data acquisition at the Division of Cardiovascular Medicine at the Radcliffe Department of Medicine, we know that the low resolution images were acquired by sampling only a portion of the k-space, masked by a rect function. Image space and k-space are connected via the Fourier transformation. The Fourier transform of the rect function is a sinc function, which contains rippling. This is the reason for the ripple artifacts appearing in the gel, discernible with the naked eye in the low resolution acquisitions. In order to generate synthetic low resolution views, with perfect alignment, we operated the down-sampling element of the acquisition model in k-space after rotation of the 3D image. This acquisition model works in a multi-resolution frame, as the high resolution volumes are arrays of size $384 \times 384 \times 384$, whereas the low resolution images are 128×128.

Fig. 4.7 shows a cut through the high resolution volume (bottom left), a true low resolution acquisition (bottom middle), and a synthetic low resolution acquisition (bottom right). The top left, middle, and right show corresponding k-space magnitude images. The low resolution 2D images represent the average of a small stack of high resolution slices, whereas we show the middle slice of the stack for the high resolution example. This is the reason why some of the structures look different between the high and the low resolution images.

Acquisition model used in the algorithm in Image space

Our method does not aim to recover the original $384 \times 384 \times 384$ volumes that were acquired for us. We have designed a second acquisition model used in the regularisation algorithm. It slightly differs from the above pre-clinical acquisition

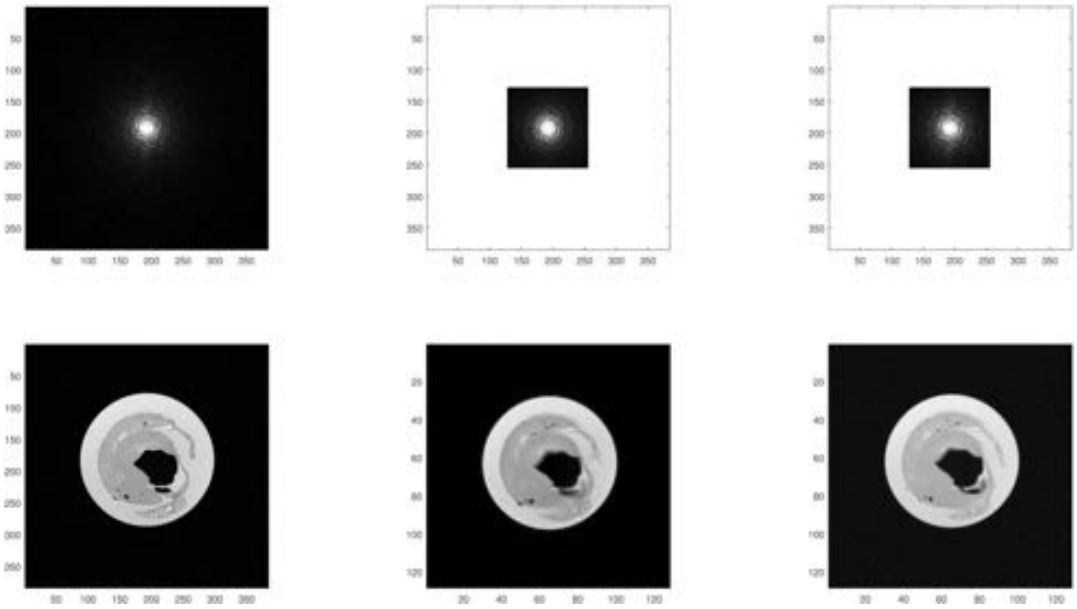


Figure 4.7: Top row shows k-space corresponding to the image space data of the bottom row. From left to right: High resolution acquisition, low resolution acquisition, synthetic low resolution acquisition.

model, which reduces the in-plane size from 384×384 to 128×128 pixels by k-space truncation. Instead, this acquisition model only down-samples in the slice selection direction. We follow models proposed in the literature such as [47] including geometric transformations, convolution with a Gaussian kernel point-spread function, and a downsampling operation. We rotate the 3D volume, to extract a slab with size corresponding to the slice thickness and down-sample it with Gaussian averaging in the through-plane direction. An example of synthetically generated data from a clinical 3D scan can be seen in Fig. 4.6, where intensity differences between short and long axis are visible, e.g. the blood pool has higher image intensities in the short axis than the long axis.

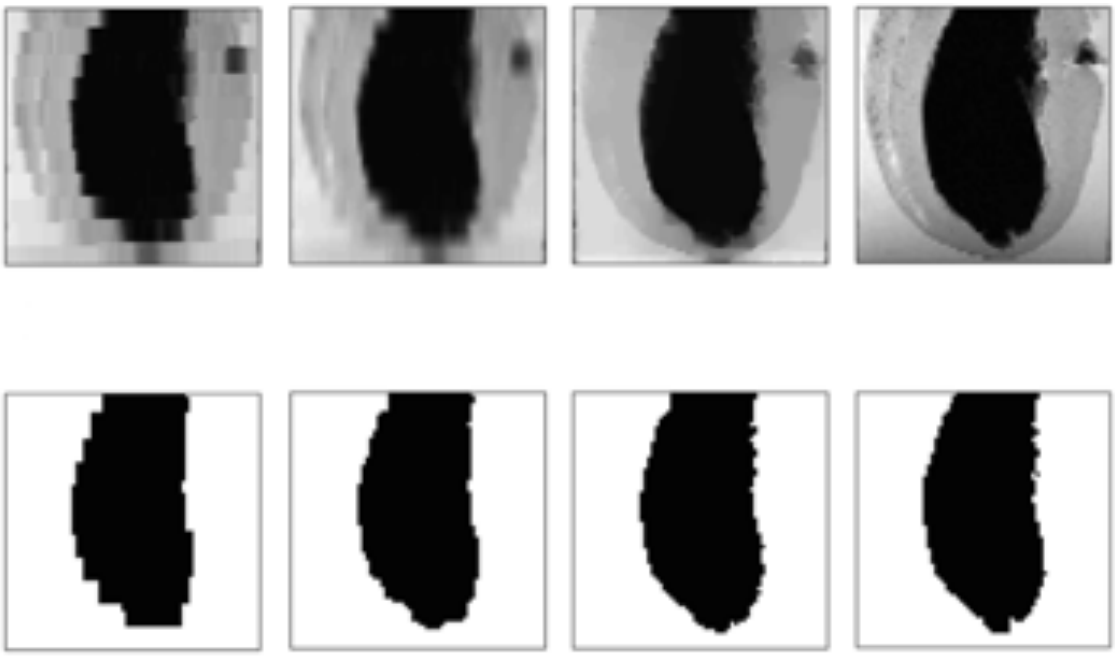


Figure 4.8: Result using 11 short axis and 11 long axis slices.The top row shows reconstructions and the bottom row segmentations achieved by thresholding. From left to right: nearest neighbour interpolation, trilinear interpolation (both interpolations using short axis data only), proposed result, ground truth. Adapted from [9].

4.5 Results

4.5.1 Pre-clinical acquisitions

The code was written and run in MATLAB R2015b. 500 iterations were sufficient for convergence, requiring a running time between 2.5h and 3h (depending on the slice protocol and number of slices chosen) on a 3.5GHz Intel Xenon CPU. Fig. 4.8 shows the result on a rat heart using 11 short axis and 11 long axis slices, and includes nearest neighbour, trilinear interpolations, our result and the ground truth. The images are cuts through the volume in a long axis orientation taken at a view that was not covered by an acquisition. The bottom row shows segmentations obtained by thresholding. Fig. 4.9 has the same format as Fig. 4.8 and shows the result with 6 short axis and 6 long axis slices, which was the combination that

ended up giving the best results.

	9 SA, 3 LA	**6 SA, 6 LA**	**3 SA, 9 LA**
PSNR (Interpolation)	18.9dB	18.18dB	15.43dB
PSNR (Reconstruction)	18.38dB	19.76dB	18.23dB
SSIM (Interpolation)	0.85	0.83	0.67
SSIM (Reconstruction)	0.82	0.88	0.78
VM (Interpolation)	3.81%	4.42%	7.72%
VM (Reconstruction)	4.49%	3.72%	6.25%

Table 4.1: Quantitative evaluation of the reconstructions of pre-clinical data. PSNR, SSIM, and Voxel Misclassifications (VM) for interpolation and reconstructions of the different slice combinations totalling 12 slices.

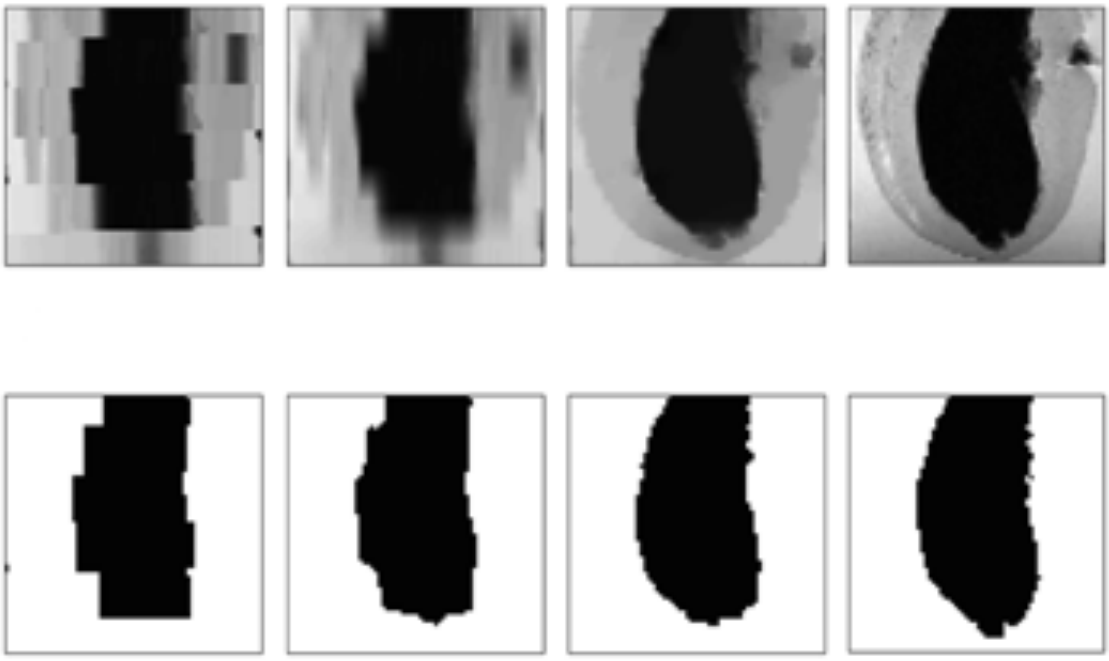


Figure 4.9: Result using 6 short axis and 6 long axis slices. The top row shows reconstructions and the bottom row segmentations achieved by thresholding. From left to right: nearest neighbour interpolation, trilinear interpolation (both interpolations using short axis data only), proposed result, ground truth. Adapted from [9].

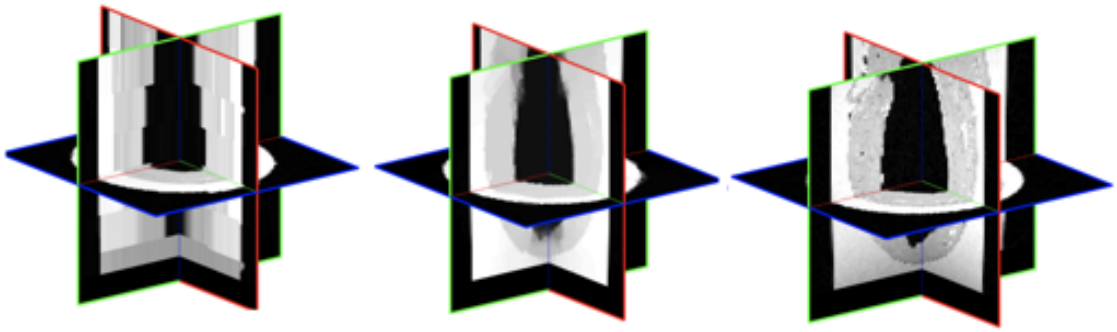


Figure 4.10: Result using 6 short axis and 6 long axis slices with the bottom short axis slice missing the apex. 3D slice display. From left to right: nearest neighbour interpolation of short axis slices, our result, ground truth.

The results are three-dimensional, therefore we aim to show 3D representations in addition to 2D slices taken from the 3D volume. Fig. 4.10 shows a nearest neighbour interpolation, the proposed result, and ground truth in 3D by showing 3 orthogonal slices through the volume, from another pre-clinical volume. In this dataset we have chosen to shift the short axis acquisitions such that the bottom acquisition does not capture the apex, which can be seen in the leftmost volume and is a common issue in clinical acquisitions. The reconstruction, in the middle of Fig. 4.10 shows a reconstructed apex. Numerical results are shown in Table 4.1, which contains average PSNR, SSIM and voxel misclassifications for ten reconstructed volumes in three different configurations of balanced short and long axis slice numbers, comparing interpolation and our results to ground truth volumes.

4.5.2 Clinical acquisitions

Fig. 4.11 shows 3D views of results on clinical data from synthetically generated low resolution multi-slice 2D data, using 6 short axis and 6 long axis scans. We show nearest neighbour interpolation, our result, and the ground truth. Fig. 4.12 shows results in a 2D view cut through the volumes, from another patient here we

show trilinear interpolation, our result, and the ground truth. Table 4.2 displays the results from quantitative evaluation of all datasets using the PSNR and SSIM..

	Interpolation	Reconstruction
PSNR	29.85dB	31.03
SSIM	0.77	0.84

Table 4.2: Quantitative evaluation of the reconstructions of clinical volumes using 6 short axis and 6 long axis slices. PSNR and SSIM for interpolation and reconstructions against ground truth volumes.

4.6 Discussion and Conclusions

Here we have presented a method for super resolution reconstruction of multi-slice cardiac MRI in 3D. We show a MAP approach making use of directional total variation to guide reconstructions from short axis and long axis slices in a contrast

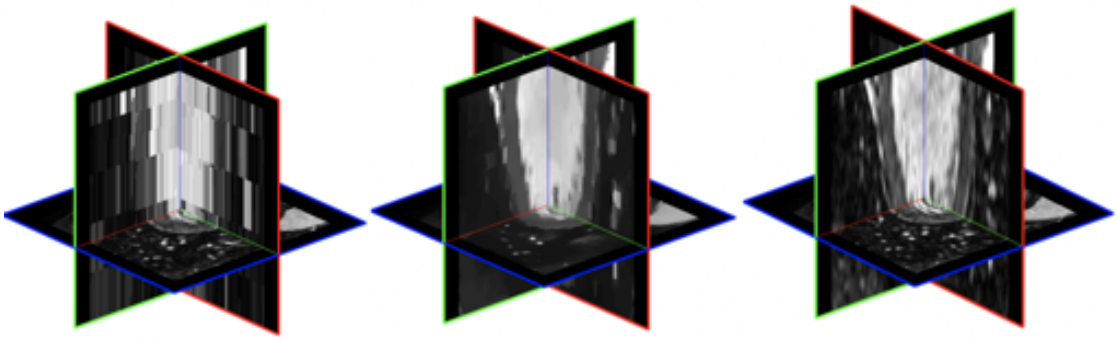


Figure 4.11: Result on synthetic clinical data, using 6 short axis and 6 long axis slices. From left to right: 3D views of nearest neighbour interpolation of the short axis images, the proposed result, and the ground truth volume. This is patient number 17 from the HVSMR dataset.



Figure 4.12: 2D cuts through volumes of synthetic clinical data, using 6 short axis and 6 long axis slices. From left to right: trilinear interpolation (initialisation using short axis data only), proposed result, ground truth. This is patient number 18 from the HVSMR dataset.

independent manner, validate it on specially acquired pre-clinical ex vivo data and test it on in vivo data.

After reviewing the literature we came to the conclusion that there are two realities regarding clinical practice and methods presented in the literature when it comes to cardiac MRI super resolution reconstruction techniques. There exist the recommended clinical imaging protocols which are almost universally used in practice, and there are super resolution reconstruction studies that use scanning protocols different from these standards without commenting on this disparity. Instead, the data used in these studies is optimised for reconstruction purposes and maximise overlap by increasing the number of total images acquired, and therefore scanner time. The number of acquisitions are often increased by a factor of two or three, while standardised protocols have been designed to limit scanner usage time. Unlike a number of recently learning based proposed methods [74, 84, 88], the proposed algorithm does not incorporate any prior information based on learning. (More on learning based methods in chapters 5, 6, and 7.) We present an alternative to the impractical multi-stack protocols followed by several studies in the literature. We have aimed to do so by trying to reconcile some of the underlying arguments from standardised protocols with the argument that increased overlap between the

slices is favourable to image reconstruction algorithms. One of our contributions lies in the proposition and the demonstration of the efficiency of a protocol comprising a balance between slices in a short axis stack of parallel acquisitions and a number of rotated long axis images orthogonal to the short axis stack, and spreading out like a fan. All of the long axis slices contain landmark features such as the apex, base and extensive cover of the boundary between the blood pool and myocardium. This protocol displays excellent coverage and overlap at the apex and the base, by choosing the centre of rotation for the long axis slices to be approximately a line connecting the apex to the base.

In Fig. 4.8 we show our result on the acquired 11 short and 11 long axis slices and compare it to interpolations and the ground truth: in addition to that we show segmentations. Both qualitatively and quantitatively, our proposed method outperforms the interpolations. All the planes we chose to display were taken at an angle that was not covered by the original long axis data used in the reconstructions, to show the qualitative results more significantly. We show results using a balanced total number of slices in Fig. 4.9 and all the subsequent figures. In Fig. 4.9, we show a 2D long axis view of the volumes, while in Fig. 4.10 we show a 3D object to contextualise the results and show the recovery of the apex. We have left the apex out on purpose in this particular synthetic acquisition, in order to display the ability to incorporate information from the long axis, when the short axis coverage falls short.

The quantitative results on the pre-clinical data are summarised in Table 4.1, which contains average values of PSNR, SSIM, and voxel misclassification percentages. The directional total variation allowed to incorporate structural information from long axis data to the short axis slices, even though both had differing contrasts. While the reconstructed images are fairly smooth, owing to total variation, the resulting images manage to reconstruct a structurally more accurate image. This means there is some trade-off with directional total variation where there are structural

improvements at the expense of texture and fine detail. If we were not using directional total variation, we would not be able to incorporate long and short axis data displaying intensity differences into the same reconstruction using a regularised least squares approach. The results for PSNR, SSIM, and voxel misclassifications show the best performance is obtained when using an equal number of short and long axis slices. It is interesting to note that for the 9 / 3 combination, the interpolations seem to give better quantitative results. We believe that of the reasons for this is the somewhat overly smooth appearance of our reconstructions. Another reason we may think is responsible for this, is the bias field artifact present in all the pre-clinical high resolution scans. It can be clearly seen in Fig. 4.10 where the bottom slice on the left hand side volume is significantly darker than the other ones, the bias field is also visible in the ground truth volume on the right hand side: there is a gradual change in darkening intensity at the bottom of the volume. This artifact is not present in our reconstruction, seen in the middle volume. The bias field is also visible in the interpolations of Figs. 4.8 and 4.9, at the bottom of the right hand side volume of Fig. 4.5, and both objects in Fig. 4.4. However, the artifact is noticeably absent in our reconstructions. For the 3 / 9 combination results, it makes sense that the interpolations are worse since there is much less short axis data present for the interpolations. It is worth noticing that the results using 9 long axis for guiding the reconstructions are slightly worse than the one using 3 long axis and 9 short axis ones. These results somewhat confirm the validity of clinical acquisition protocols, which have not been designed for maximisation of overlap and with super resolution reconstructions in mind. The best performance in our experiments, however comes from using 6 short and 6 long axis slices. Balancing the 12 acquisitions to 6 short and 6 long axis slices shows improvements which outperform the interpolations, as well as the reconstructions with other combinations of slice numbers. This highlights our finding that the slice protocol followed in clinical practice is not ideal for 3D reconstruction, and that increasing the number of long axis slices improves the reconstruction. We show that when applying post-processing to reconstructions, a balanced number of slices seems to be the best way of using a set number of

acquisitions. Based on this result, we chose to keep a balanced number of slices for the evaluation on in vivo data. We show a set of volumetric plots in Fig. 4.11 and a 2D plots in Fig. 4.12, on a different dataset. Both reconstructions show an improved reconstruction with smoother myocardium boundaries. Quantitative results are shown in Table 4.2, where the best results are highlighted in red.

In this work, we have addressed the problem of combining structural information from long axis images to improve 3D reconstructions from short axis images. Accurate 3D volumes are required for the generation of meshes for electromechanical models as well as other applications such as measuring cardiac volumes or estimating ejection fractions. Advantages of our method include the ability to reconstruct volumes from one acquisition without the need for extensive training or databases, as well as the ability to treat short and long axis slices in a contrast independent manner. Another advantage over some methods in the literature, is that our method does not require additional slices, merely a different allocation of the number of slices per view to get the best performance. There are some limitations to this study, starting with long computation times. In order to overcome this issue, the code can be optimised, parallelised and run on a GPU instead of just the CPU as we have done. The most computationally expensive part is the 3D to 2D sampling using the acquisition model during the data consistency check. The algorithm assumes a preprocessing step of registration between short and long axis slices, and any inaccuracies in this step will be propagated into the image reconstruction. Furthermore, as it is based on total variation, regions outside the sampled planes will typically be as smooth as possible. It is therefore important that as much of the heart as possible is imaged by at least one plane, which is not the case in current clinical multi-slice cardiac MRI where for example the majority of the apex is often left out. Between the spokes of the long axis slices, only the short axis is driving the reconstruction, and is highly sensitive to initialisation. In order to translate this contrast independent method to clinical multi-slice data, additional pre-processing to minimise slice misalignment would be required as well as a change in the slice

protocol to maximise slice overlap. Our results suggest some recommendations for protocols that depart from current clinical practice, in particular that one should balance the number on short axis and long axis slices in the acquisitions. Our results also suggest that the long axis slices, as defined in current protocols, are suboptimal for 3D reconstruction. Maybe clinical protocols should consider oblique planes, something we have not explored in our balanced slice combinations. The downside of such oblique slice recommendations is the difficulty of interpreting these non-traditional views by clinicians.

In this chapter, we have presented our work on a *traditional* approach to super resolution reconstruction of cardiac MRI including short and long axis data. Future steps to this could include incorporation of temporal information. An interesting further step to this work would be to acquire real in vivo multi-slice 2D data using an equal number of short and long axis slices, which would require some pre-processing to correct for misalignment [89]. With the field rapidly moving towards another direction, we decided it would be more productive to explore deep learning approaches. In the next chapter, we present background review and relevant literature on deep learning methods applied to super resolution of cardiac MRI, in order to introduce the remaining two analysis chapters presented in this thesis.

5

Background on Deep Learning and Literature Review on its application to Super Resolution

The previous chapter dealt with a novel approach to reconstruction using regularisation of minimisation of least square differences regularised by three-dimensional directional total variation. On one hand, the drawbacks of such methods include the long computational times needed for every dataset one wants to reconstruct as well as the difficulty to generalise to highly variable examples, on the other hand one of their main advantages is the freedom from the necessity of large datasets or a learning process. The increasing availability of large labeled datasets in medical imaging and of computational power has given rise to learning based techniques. In recent years, a majority of research in the field of computer vision and imaging has been devoted deep learning and especially Convolutional Neural Networks. Once the learning stage is completed, such methods can perform their tasks almost in real time on a single example. This chapter deals with key concepts of deep learning methodologies and gives the reader a review of the literature relevant to the use of deep learning in super resolution and medical image processing.

5.1 Deep Learning

Deep learning refers to a class of techniques in machine learning, which allows models to learn representations of data and to perform a plethora of tasks. Deep learning makes use of example data to learn complex underlying patterns via carefully designed models consisting of varying number of layers, also referred to as neural networks, and training regimes. Deep learning has allowed to advance state of the arts results in many different fields such as computer vision, natural speech processing, as well as medical image analysis [90]. Neural networks, as the name may give away, have somewhat been inspired by the behaviour and structure of neurons and axons in the brain. Neural networks are typically made up of connected nodes. Such networks are also referred to as feed forward networks, because the data travels from the input to the output channels via the nodes making up the layers in between. A simple neural network can be seen in Fig. 5.1[1]. Hidden layers is the name given to the layers which exclude the input and output layers. An input layer and an output layer are connected by a number of hidden layers, in this example there are two hidden layers. Here the first hidden layer has 4 nodes and the second hidden layer has 6 nodes.

Networks are characterised by the weights and their architecture. The weights are the trainable parameters, and the architecture is what defines the structure of a network, including the number of hidden layers and nodes. Layers are made up of nodes, which are connected to nodes from adjacent layers. There are 4 nodes in the first hidden layer, which are each connected to the 6 nodes from the second hidden layer in Fig. 5.1. Every node sends its output to every node in the layer below, this connectivity is referred to as *fully connected*. A node receives input from adjacent nodes upstream (or the input layer if the node is in the first hidden layer) and generates an output which is sent to adjacent nodes downstream (or the output layer if the node is in the last hidden layer). Every input to the node has a weight $\mathbf{w}$ which determines how much the input contributes to the output of that node, and

[1] Figure generated using `https://gist.github.com/craffel/2d727968c3aaebd10359`

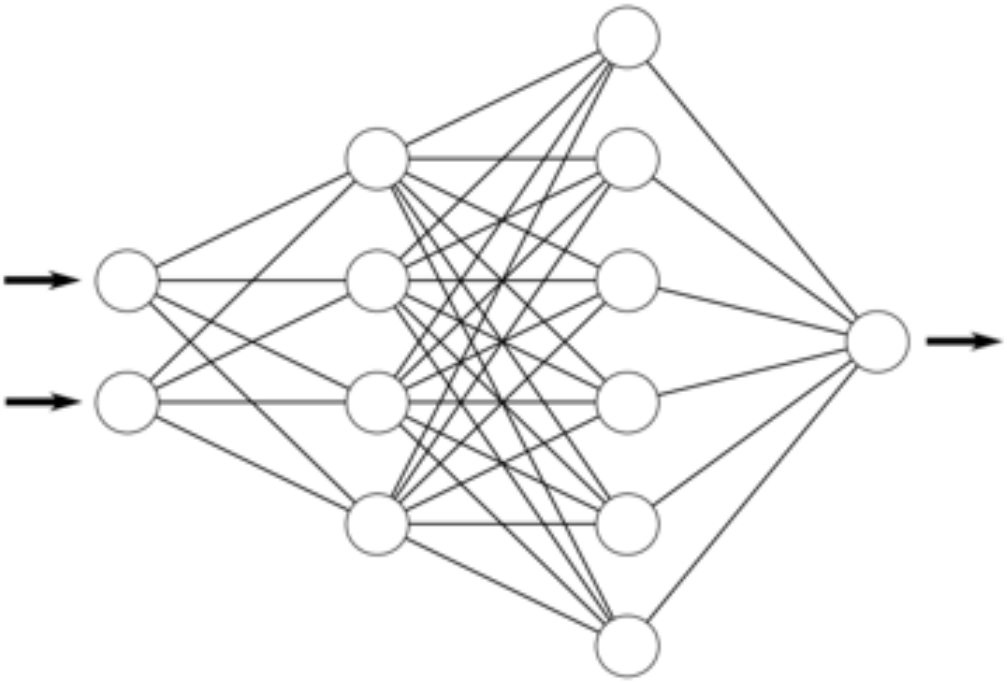


Figure 5.1: A neural network with two hidden layers.

every connection connecting two nodes has its own associated weight. To calculate the output of a j^{th} node in a layer, a common operation is the weighted sum of the input to the node $z_i = \Sigma w_{ij} x_i$. A non-linear function f is applied to z_i to obtain the output of a node $y_i = f(z_i)$ which is the input to the subsequent node [91]. Linear models allow to separate features in feature space through straight lines while introducing non-linearity elevates this concept and enables to draw various shapes for boundaries on feature hyperplanes. Successive linear functions could be formulated as a single linear function and do not add to the representational capacity of a network. It is the non-linear functions which allow neural networks to represent data in such complex ways. These non-linear functions are also referred to as the activation functions. Examples include the sigmoid, hyperbolic tangent, and the rectified linear unit (ReLU) functions. These examples of activation functions can be seen in Fig. 5.2. The ReLU, which is a simple thresholding function setting negative values to zero, is the most popular activation function right now. It has been shown that the usage of ReLU improves learning when it is chosen as the

activation function for the hidden layers and the number of hidden layers is large (≥ 3) as shown by Glorot *et al.* [92]. In the study, the authors show the superiority of ReLU to other activation functions, and a parallel between the activation behaviour of biological neurons and ReLU is drawn since they operate in similar fashion and only fire if a threshold is reached. The fact that all negative values produce the same output could be a problem, and it makes gradients in that zone equal to zero which led to the introduction of the so called leaky ReLU [93]. Leaky ReLU are the same as ReLU activations, just that the negative elements are not discarded straight away. Instead, a small positive gradient ($y = 0.01x$) is given to negative values, this is to address the so-called dying neuron issue. Another tool used in deep learning to overcome some issues in training is called dropout. Dropout layers refers to layers that are randomly switched off at the training stage, which avoids that the network becomes too specific to the training data (an effect known as overfitting) by enforcing several different neuron combinations to be able to handle representations of the data [94].

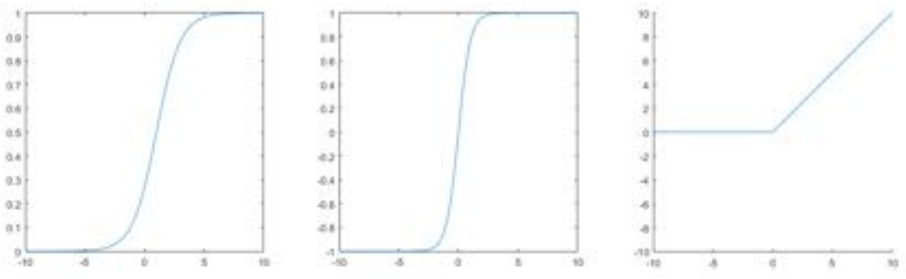


Figure 5.2: Typical activation functions: sigmoid, hyperbolic tangent, ReLU.

Training a network refers to determining the weights giving the best results whilst keeping the network architecture unchanged. The term *deep* in deep learning refers to the amount of hidden layers, the more hidden layers there are in a network, the deeper it is said to be. The paradigm of neural networks is to subdivide complex

tasks into several simpler functions. The more complex an architecture is, the more complex representations of data it can achieve and tasks it can be trained to perform. However, the deeper an architecture is, the more the number of parameters may increase which also amplifies the requirements for training data, and increases the risk of overfitting. There exist several forms of learning. The most common form of learning is supervised learning, where all of the training data has associated labels. Other learning strategies include unsupervised, semi-supervised and reinforcement learning. Unsupervised learning uses data which has no labels, while semi-supervised learning uses labeled as well as unlabeled data in its training process, assuming an underlying distribution of the unlabeled data. Reinforcement learning works by using a reward system depending on the output, which is suitable when the task at hand has a reward system, such as a game with a winner and a loser. One such example is AlphaGo [95], a programme which has received extensive media coverage for being the first machine being able to beat professional players at the game of Go.

Supervised learning refers to a training regime for models where input data and output ground truth (which can be labels, images, heat maps, etc.) are available for training. In a medical setting, data may be in the form of images or patient data such as heart rate readings, and the labels may be the segmentation of an organ in the image or a predicted diagnosis for the patient. The model produces a prediction of the label based on the input data, for example: it can be a class or set of classes in a classification problem, or a binary pixel value in a segmentation task. At the beginning of the training task, all the filters are generally automatically set to random values. In the training stage, the data is fed to the model, which produces an output label. Objective functions are used in optimisation problems, they map out real values representing an error or reward, as a function of the variables in a multi-dimensional feature space. In the case of deep learning, the variables are the weights, meaning that the problem is highly multi-dimensional since the number of weights often easily go into the millions. Objective functions are often loss (or

error) functions, making the aim of the optimisation to minimise the error of the output. Such a loss function computes the error between the predicted result and the desired result and the model adjusts its parameters, the weights, to reduce the output error [91].

Gradients are used to determine how the error would change for each adjustment. A negative gradient pointing in the direction of steepest descent would guide the network to a minimum where the error would be lower than at the current configuration [91]. A popular procedure is Stochastic Gradient Descent (SGD), which works by taking small selections of random (stochastic) samples, also referred to as batches, and adjusting the weights according to their gradients. The process is repeated until the value of the loss function stops decreasing, indicating a minimum has been reached. Minimising a loss function f, can be seen as minimising a function having the form of a summation $f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w})$ where f_i is the result of the function being evaluated at the i^{th} of n data points using the parameters, or weights $\mathbf{w}$. In SGD the gradient is approximated for each datapoint. The method is summarised in Algorithm 1. At every iteration, the gradients of the objective function are computed with respect to the weights, and the difference is taken to get an update on the weights. In real practice, the gradients are often too large, so in order to balance out the contribution and avoid over-correction, the gradients are scaled by the so-called step-size or learning rate, α, before performing the weight update step.

Algorithm 1 Stochastic Gradient Descend

α = Learning rate,
s = batch size
$\mathbf{w}$ = weights
while $\mathbf{w}_t$ not converged **do**
    $g_t \leftarrow \frac{1}{s}\sum \nabla f_i(\mathbf{w_t})$ (Get gradients)
    $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \alpha \cdot g_{t-1}$ (update weights)
end while

Another optimiser that has gained popularity recently is the so-called Adam optimiser. The name is derived from Adaptive moment estimation [96]. It is the optimiser we have used in all our experiments described in chapters 6 and 7. Some of the main differences to SGD are the incorporation of momentum and adaptive learning rates. SGD makes constant steps towards a minimum, without taking into account any of the previous steps that have been taken. Momentum helps avoiding getting stuck in local minima, analogous to a ball rolling down a hill having momentum and being able to roll upward if needed over an insignificant bump. Momentum is a way of giving gradient descent the ability to incorporate information on prior steps and use it in the update. Instead of monotonous steps $\mathbf{w}_t = \mathbf{w}_{t-1} - \alpha \cdot g_{t-1}$, the moment of prior steps is incorporated as such: $\mathbf{w}_t = \mathbf{w}_{t-1} - \alpha \cdot g_t + \beta(\mathbf{w}_{t-1} - \mathbf{w}_{t-2})$ where $\beta \in [0, 1]$. This results in an exponential decay of the contribution of previous gradients, and the standard formulation is recovered when $\beta = 0$ [97]. Adam optimisation pseudocode is shown in Algorithm 2, and the authors of [96] also include a recommendation on improving the efficiency of the algorithm related to the adaptive learning rate $\alpha_t = \alpha \cdot \sqrt{1 - \beta_1^t}$ and $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \alpha_t \cdot \frac{m_t}{\sqrt{v_t} + \hat{\epsilon}}$, which allows the learning rates to adapt for different weights using first (m_t) and second (v_t) moments.

Algorithm 2 Adam

α = Learning rate,
$f(\mathbf{w})$ = Objective function,
n = Number of samples,
β_1, β_2 = First and second moment decay rates,
ϵ= A constant to avoid divide by zero
$\mathbf{w}$ = weights
while $\mathbf{w}_t$ not converged **do**
  $t \leftarrow t + 1$
  $g_t \leftarrow \frac{1}{n} \nabla_{\mathbf{w}} f_t(\mathbf{w}_{t-1})$ (Gradients)
  $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (first moment)
  $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (second moment)
  $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (bias corrected first moment estimate)
  $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$(bias corrected second moment estimate)
  $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t + \epsilon})$ (update weights)
end while

Training is completed either when convergence has been reached or when a set number of epochs have been completed. An epoch is when all of the training data has been seen by the network one time, in a forward and backward pass through the network (more on this in the subsection below on Backpropagation). If we have 100 training samples, and a batch size of 10, it takes 10 iterations to complete an epoch. Higher batch sizes will require larger amounts of memory. Once training is completed, the performance of the model is tested using testing data that has been set aside specifically for testing. It is important to keep aside a portion of the entire data just for testing, so that the capacity of the model to generalise may be evaluated on unseen data. In addition to setting aside testing data, another part of the data is kept aside for validation. The validation dataset can be used to estimate hyperparameters such as the number of epochs and to fine tune the architecture of the network and determine network performance during the training stage by looking at the intermediate results on the validation set after every iteration. Selecting the right architecture is the only part that requires the real fine tuning and user input. Some of the hyperparameters such as learning rate and numbers of filters are also generally fine tuned by users empirically. Weights are adjusted, or learnt, automatically during the training process aided by a method called backpropagation.

Backpropagation

The gradients used in optimisation algorithms in deep learning are evaluated using backpropagation. It is called backpropagation because the gradients of the objective function are evaluated with respect to the input of the model, or a hidden unit, by going backwards through the network and using the gradient with respect to the output of the previous hidden unit, or input, using the chain rule. At the start of the backward movement in the backpropagation, the final outputs of the model are compared to the ground truth labels to obtain error derivatives [98], and the errors are propagated backwards through the model to determine gradients further

upstream in the architecture by LeCun et al [91]. One of the first successful uses of backpropagation was described by the same author in [99], where a special kind of neural network was used for recognition of handwritten digits. This study made use of Convolutional Neural Networks (CNNs). So far, we have focused on deep learning without mentioning CNNs, which are at the centre of the next two analysis chapters. CNNs are neural networks particularly well suited for working with array data which includes image and video data.

Batch normalisation

While networks as a whole learn representations of the input data statistics and map it to the statistics of the output, each single layer on its own relies on previous layers and subsequent layers depend on that layer too. The distributions of the inputs to every layer change in the training process with the changes of previous layers, which makes the whole training process more complicated. The authors of [100] refer to this as internal covariate shift and propose batch normalisation to overcome the aforementioned difficulties. This consists in normalising every batch whenever a batch normalisation layer is implemented, which reduces the amount of shifts in data distributions by ensuring the mean and variance remain the same and makes training more robust and efficient. Instead of having to cope with changes in the statistics of adjacent layers, it allows to learn more efficient mappings and enables higher learning rates to be used resulting in faster convergence. As a result, the authors show accuracies increasing and drastic reduction in number of steps required to reach a certain accuracy threshold when training state of the art image classification models. Batch normalisation is inspired by input normalisation, which is shown to vastly improve training performance in [101], and taking it a step further by applying it inside the network as opposed to just a data pre-processing step. In the same paper, the authors give advice on other practices for deep learning such as shuffling training data. Shuffling will separate successive data points that

may be too similar to one another, and produce small errors. This correlates with the aforementioned stochastic aspect of SGD.

5.2 Convolutional Neural Networks

5.2.1 Background

CNNs are an extension of neural networks and are able to efficiently handle data in arrays, such as images and video. A typical image with three colour channels would be a $h \times w \times 3$ array, a typical cardiac cine MRI sequence an array of $128 \times 128 \times 30$, whereas a typical frame of a cardiac cine MRI sequence would be a 128×128 array (for convenience, we will refer to an array of this size as $\mathbf{A}$ from here on.) Neural networks described above would not be able to handle data this large, vectorising the small grayscale 2D image $\mathbf{A}$ pixels would result in an input vector of length 16384, meaning that every node in a layer would increase the number of weights by that amount because of the fully connected connections. Taking one hidden layer of the same size would result in over 268 million parameters. It is worth noting that many CNNs deal with much larger image sizes, which would carry even larger parameter numbers in fully connected neural networks. Instead of being fully connected, CNNs form local connections by using successive layers made up of several small sized kernels, which drastically decreases the number of parameters. Instead of requiring 16384 weights, a simple 3×3 kernel would correspond to 9 parameters and can be swept across the whole image to generate a feature map. This means that all of the pixels (we will be talking about the application to images so a datapoint is a pixel) in one layer do not connect to all of the pixels in the next layer, instead small regions connect to small regions in adjacent layers. Stacking and increasing the number of layers allows to effectively extend small local connections to a larger field of view.

CNNs are very similar to the neural networks discussed in the above section, with the differences being that convolutional layers replace the nodes where values of the convolutional kernels or filters effectively replace the weights, and the addition

of maximum pooling layers. One node may be seen as being replaced by one convolutional filter. CNNs also involve fine tuning of architectures and learning of parameters detecting statistical patterns in the data, aided by non-linearity of the model introduced in part by activation functions. Convolutional layers are made up of collections of filters, or filter banks. The term convolution comes from the mathematical operation used to apply a filter to an array. Filter banks may be made up of many filters, allowing to extract a wide range of features. A filter may be a 7×7 kernel, the application of one filter to an array once results in a feature map, a different representation of that image. If we take the first layer to have a filter number $n = 32$, that results in a filter bank of size $7 \times 7 \times 32$. Passing the image $\mathbf{A}$ through this filter bank would result in an output vector of feature maps of size $122 \times 122 \times 32$. The resulting feature maps decrease in size (from 128 to 122 in this case) compared to the input image because of the convolution, unless padding is applied. Here the stride is 1, if different values of strides are used, the output will shrink even more.

In CNNs, the learning process consists in learning values of the convolution filters which will best highlight features activating neurons. Training is done using backpropagation and optimisers such as SGD and Adam as above. In an image represented by an array of numbers, features are reflected in collections of local values which can be easily highlighted through filtering. One of the strengths of CNNs is that they are able to cope with position invariant image features by learning convolution kernels. This means if a feature such as an edge is present anywhere in the image, a filter designed to highlight edges will detect it at every location where it occurs. Instead of learning pixel-to-pixel dependencies, a CNN learns mappings between collections of pixels. The range of features increases with successive convolutions. Stacking two convolutions with kernel size 3×3 and stride 1 would result in an effective receptive field size of 5×5. In this way information gets propagated through the network. CNNs work in a hierarchical way, where the first layers highlight simple low level features such as edges, which make up

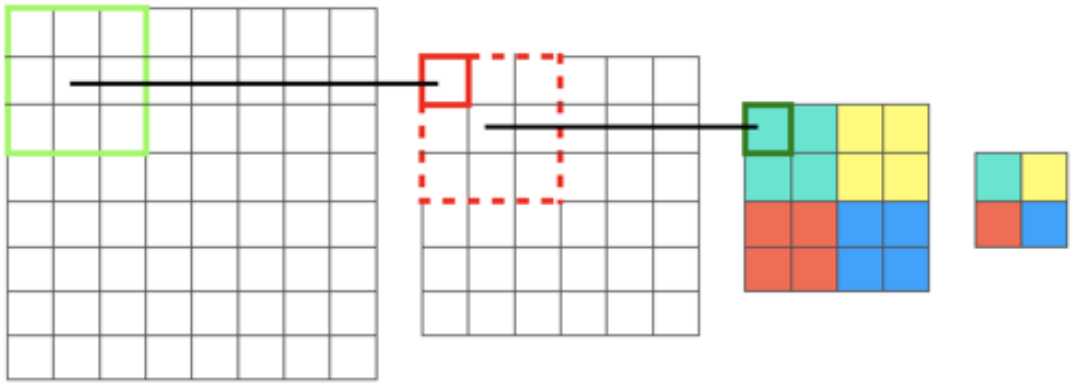


Figure 5.3: A visualisation of a two successive 3×3 convolution operations of stride 1 followed by a 2×2 pooling operation.

higher level features such as objects appearing through repeated applications of convolution. When looking at the output at the first layer often various edge finding operations are revealed by the feature maps. Successively applying convolutions increases the effective receptive field and is one of the strengths of CNNs. A low level edge may be making up part of a whisker, which is making up part of a cat face [102]. The deeper a CNN goes, the more complex features may be extracted. Successive convolutions only allow linear extensions of the receptive field size, thus in practice CNN architectures include down-sampling operations to further increase the receptive field size. Maximum pooling is such a down-sampling operation, where only the maximum value of the selected pool is carried forward. Fig. 5.3 shows two successive convolutions of a 8×8 array with a 3×3 filter, with a stride of 1, followed by a pooling operation, resulting in a 2×2 output. If this was a maximum pooling operation, the maximum values of each quadrant represented by the different colours would be carried towards the output, which adds further non-linearity. An

alternative to maximum pooling is average pooling, where the average of the values is carried over instead of the maximum. Maximum pooling highlights features such as edges whereas average pooling smoothes those features. Down-sampling can also be achieved by increasing the strides of the convolution filters.

5.2.2 The rise in popularity of Convolutional Neural Networks

While CNNs were around in the 90s, they have only started taking over in recent years. The origin of the renewed interest in CNNs may be traced back to 2012 where the ImageNet Large Scale Visual Recognition Competition (ILSVRC) recognition challenge was won by Krizhevsky *et al.* [103] and their so-called AlexNet network, a CNN that gave an error rate that was more than ten percentage points ahead of the runner up entry. The authors used a Graphics Processing Unit (GPU) to train a network on a large dataset of 1.3 million images. The availability of large datasets and computing power rekindled massive interest and has led to an explosion of the field. The architecture of AlexNet is made up of 5 successive convolutional layers, followed by three fully connected layers. The two parts of the network could be bundled into two modules: feature extraction and classification. The kernel sizes are 11,5,3,3,3. The fully connected layers are located at the end of the architecture when the inputs have significantly shrunk in size and carry only 4096 weights each. A total of 60M parameters are present in the architecture. The input is an image of size 224×224 while the output is a list of the five objects the network estimated to be the most likely present in the input image and an associated probability percentage. Authors use maximum pooling, and ReLU as the activation function. They also show sixfold improvements in training speed with SGD over the use of hyperbolic tangent function. Training took 5-6 days on a GPU, and authors suggest results would improve with increased computational capacity and even more data. Another technique to which CNNs and deep learning owe their recent surge to is data augmentation. This consists in creating additional training data via operations such as rotations, flipping, and deformations. In many cases for

example, a natural image would have the same interpretability as its mirror image. Similarly, in a set of medical image data containing cells and their segmentation, rotating the image- segmentation pair in the same manner would result in a new image-segmentation pair. Apart from being necessary when data size is not large enough, augmentation also contributes to avoid overfitting. Transfer learning is another alternative method to train a network when there is not a sufficient amount of readily available data. Instead of learning networks from scratch, a network may be trained on an alternative large dataset. This model can then be used as an initialisation to the training stage to fine tune the model to the wanted application using the small amount of non generic data which is available, instead of random initialisation of the weights. This is possible because it appears that the filters in the early hidden layers display the same behaviour across different models [104].

Interestingly enough, an abridged history and evolution of CNNs may be very well followed by looking at winning (or high performing) entries to the ILSVRC for a few years following the 2012 competition being the first with a CNN winning entry [103]. In the early days of CNNs, many of their aspects were not well understood and to this day they are still considered as black boxes which to a certain extent boiled down to sole trial and error approaches (e.g. Krizhevsky *et al.* merely hint at more data and computational power to further decrease error [103] indicating a certain lack of insight). The excellent results of CNNs despite the lack of understanding have caused astonishment but also frustration in the scientific community. The authors of [102] express their frustration with the lack of a scientifically thorough approach to CNNs, and have introduced a visualisation technique which has given insight to the way the hidden layers function and how the classifiers operate. This as well as other ground breaking novel insights have allowed CNN design to become more intelligible [105, 106]. Using these tools, the authors were able to rationally identify more efficient hyperparameters such as filter sizes and strides on the same general structure as followed by the AlexNet architecture in terms of numbers of convolutional and fully connected layers. For example, the

size of the first kernel was made smaller, so as to not skip out as much of the information contained in the original image. In the ILSVRC 2013 competition, ZFNet [102] became the winning entry with their error rate achieving a reduction of over five percentage points compared to the AlexNet entry of the previous year. One of the aspects which plays an important part in network performance and definition of the architecture, is the number of hidden layers, or depth. An entry to the 2014 ILSVRC competition was the so-called VGGNet [107], which has 19 layers including convolutional layers all of which have filter size 3×3 achieving a very large receptive field while keeping small computationally inexpensive kernels. Indeed, a filter of this size has a receptive field size of 3×3, two consecutive filters of this size have a receptive field size of 5×5, three consecutive 7×7 and so on. It may also be noted that the filter numbers increase in the later layers of the network, this increase in filter numbers is also present in the later layers of ZFNet. VGGNet placed first in the image localisation challenge and second place in image classification challenge. CNN architectures and data come in various shapes. The rule of thumb is that the deeper the network and the larger the number of filters, the more data is needed to train the network effectively. The winning entry to the 2014 ILSVRC with an error rate of 6.7% was the so-called GoogLeNet [108]. The authors show that the layers in CNNs do not have to follow a sequential pattern, a rule that had been followed previously. They show that using what they call *Inception* modules (modules that contain layers which are not stacked sequentially) actually improves performance. Inception modules allow the parallel executions of pooling and convolution operations, aided by 1×1 layers for dimensionality reduction. Effectively, it is like having a network made up of networks. While on one hand this network shows a massive increase in the number of layers with over 100 layers, on the other hand it can boast a much decreased number of parameters compared to networks such as AlexNet or VGGNet. This architecture is also the first of the list not making the use of fully connected layers, which contributes to keeping the parameter numbers lower. The 2015 ILSVRC winning entry was a network with even more layers than GoogLeNet, called ResNet [109] which achieved

an error rate of 3.6%. Besides having many layers, its novelty was to introduce the idea of residual learning, which consists in adding the output of a module of layers to the original input via a shortcut connection between the input and output of a block of layers. This effectively results in an altered version of the input. Until the appearance of residual learning, the outputs would be the result of a function $x = f(x)$. The output of a residual module is $H(x) = r(x) + x$ which results in learning what is missing of $H(x)$ by adding the residual $r(x)$. Authors show that learning a residual mapping makes it easier to learn very deep networks than it would be without residual modules. Residual networks allow faster training of much deeper networks which achieve better results.

In the text above we have explained a short history of CNNs and evolution of architectures including AlexNet, ZFNet, VGGNet, GoogLeNet and ResNet. These CNNs have been mainly applied to image classification and localisation. Through the advances in understanding of CNNs as well as increased computational power, access to data and user friendly libraries, CNNs have also made their way into many other fields of computer vision and established new state of the art results, including in the field of image super resolution.

5.2.3 Convolutional Neural Networks for Super Resolution

A very simple three-layer architecture is presented in the seminal deep learning for super resolution paper by Dong *et al.*, in which the authors present their Super Resolution Convolutional Neural Network (SRCNN) [110]. This has become a benchmark and has been used for comparison in much of the subsequent literature to this date. The authors aim to model the super resolution task like it was done with dictionary learning, meaning to extract a low resolution patch and to match it to a high resolution one using two dictionaries. This is done by a very simple architecture of only three layers. The first layer consists of a filter bank with kernels of size 5×5, the second layer of size 1×1, and the third of size 9×9. The first layer represents small patch extraction, the second a non-linear mapping, and the

third a bigger patch. It is interesting to note that atoms in learnt dictionaries can appear somewhat similar to convolution filters in CNNs. Indeed, filters learn to "see" edges, textures, specific shapes, just as atoms aim to sparsely represent a myriad of image features. Another aspect of its simplicity is the fact that the filter number is quite low and a small number of images (91 natural images in total, before division into 24,800 sub-images) was used for training. The low resolution image is up-scaled to the desired size using interpolation, so the input and output images have the same size. The network was trained with a MSE loss function, since minimising MSE results in maximising the PSNR. Its results far outperform various methods including sparse coding (dictionaries) in terms of quality and running time for up-sampling with factor 3. The authors also experimented with deeper architectures and larger kernel sizes, which they found not to necessarily lead to better results while showing an increase in parameters and decrease in training and execution speeds.

SRCNN has established its status as a landmark by being applied successfully to super resolve images and enhance further analysis where it improves results drastically, such as in face recognition [111] or cell counting [112].

In later studies, the authors of SRCNN have incorporated the use of simultaneous use of the three RGB channels to further improve on SRCNN in [113] as well as introducing the fast-SRCNN (FSRCNN) [114]. The main innovation of FSRCNN was that no pre-processing is required to get the image to the same size as the high resolution label, up-sampling is achieved by the network using a transposed convolution layer.

The first application of a very deep architecture such as VGGNet, but applied to super resolution was the Very Deep Super Resolution (VDSR) network proposed in [115]. Like SRCNN, VDSR uses up-sampled low resolution images set to the size of the sought after higher resolution output image. Just like VGGNet, VDSR

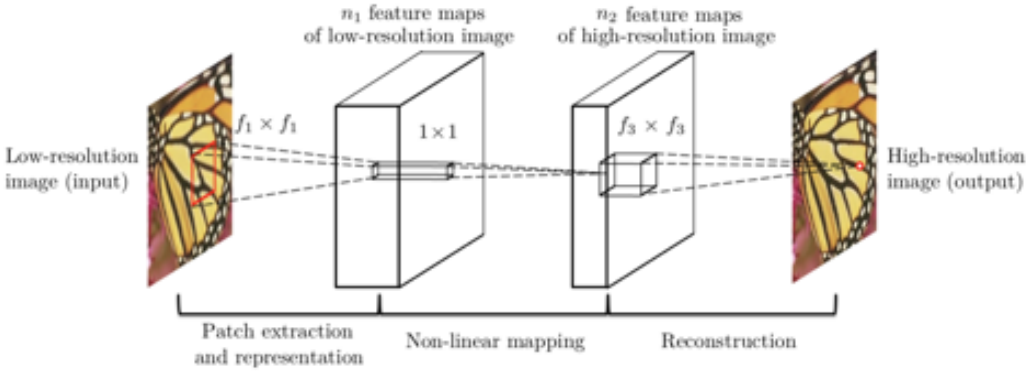


Figure 5.4: Architecture of SRCNN, showing patch extraction and reconstruction. The lightweight architecture achieves high quality results at almost real time speeds for up-sampling images by a factor of $\times 3$ [110].

consists of 3×3 convolutional kernels layered 20 times with ReLU activation function layers. The concept of residual learning proposed in ResNet [109] is used, probably owing to the depth of the architecture. This results in effectively learning the difference image between the high resolution and the interpolated images, instead of an end-to-end mapping as has been done in SRCNN [116]. Different scales of the low resolution interpolations are used to train several models of the VDSR network with all different possible permutations of training data scales. The authors performed quantitative evaluation on the ability of VDSR to enhance images by a factor of $\times 2$, $\times 3$ and $\times 4$, and compare PSNR measurements of the different models trained. The model that was trained using all the different scales of low resolution data achieves consistent quantitatively good results for all the scales.

Research contributions have evolved into super resolution strategies which do not require data pre-processing steps such as up-sampling or interpolation to match the low resolution input size to the output size by introducing up-sampling layers into the network architecture. An example of this is the improvement made from

SRCNN to FSRCNN which resulted in faster processing times. In [117], the authors propose the Gradual Up-sampling Network (GUN), which gradually up-samples low resolution images. The network is made up of 3×3 convolutional layers with padding to avoid shrinking, ReLU, and up-sampling layers which resize the input to a slightly larger size. Repetition of such layers allow the up-sampling of the smaller input images. The End-to-End Deep and Shallow Network (EEDS) [118] also does not rely on pre-processing, such as interpolation or up-sampling to the required size, meaning that the size of the input and output in EEDS are not the same unlike in SRCNN or VDSR. The architecture of EEDS processes the deep and shallow networks in parallel. The deep network contains 13 layers which make up three modules acting as feature extractors, up-sampling, and multi-scale reconstruction. The shallow network consists of only three layers having the aim to make up the same three modules as the deep network and is meant to add stability to the training of the deep network by incorporating long and short range contextual information in the images. The outputs of networks are concatenated in the final layer of EEDS.

One of the contributions of the study presented in [119] is the super resolution ResNet SRResNet, where the ResNet was pretty much just edited to minimise a MSE loss. The authors of [120] propose an Enhanced Deep Residual Super Resolution (EDSR) network in contrast to SRResNet. The EDSR network does not make use of batch normalisation, resulting in simplified residual blocks. The authors argue that batch normalisation is better suited for more abstract high level tasks such as localisation or classification, and in super resolution tasks the input is strongly correlated to the output. In addition to that, the training is accelerated and the performance of the network is shown to improve when pre-training the network is performed with training data of scale of $\times 2$ before training on scale $\times 4$, which can be seen as applying some sort of transfer learning.

All of the aforementioned CNN approaches to super resolution aim to minimise the MSE (L2 loss), or slightly different functions such as the L1 loss, which results in a

better performance when it comes to PSNR comparisons. In [121], the authors argue for use of the L1 instead of the L2 norm for neural networks in image processing, because the L2 norm amplifies penalties from large errors. CNN super resolution methods seem to have reached a plateau of quantitative improvement which is also reflected in qualitative aspects of the results. Many of the results appear overly smooth, which is a result of minimising MSE. Recent efforts were made on the aim of enhancing the perceptual quality as opposed to the standard numerical quantitative evaluation. Such an effort is shown in [122], where style transfer using perceptual loss is performed. The sought result of the transformation network is to transform the input into images with a different style, such as a pencil sketch, oil painting, or with a more realistic appearance. The authors use pre-trained classification networks, since the perceptual features are encoded in such classification networks, in addition to pixel-to-pixel loss functions such as the L2 norm and the total variation. This means their super resolution CNN has an additional network which acts as a loss function. Numerical results of quantitative evaluation do not match the visual qualitative enhancement. It can be argued that the visual improvements matter more than quantitative metrics. It can sometimes be difficult to assert the strengths of certain methods with this rapid growth that is being experienced right now, when some studies show images that look qualitatively identical and boast about very small gains in PSNR, while other studies with the opposite where no improvement is seen in terms of PSNR but the image enhancement is more visually noticeable. More in depth reviews can be found in [123], and [124], the author of the latter adds remarks to and challenges some of the results or omissions in publications, something which is important in this field at this point in time. Perceptual losses are at the core of Generative Adversarial Networks (GANs) which will be described in section 5.3.

5.2.4 Convolutional Neural Networks for Medical Image Analysis

The U-net

A prominent example of CNNs applied to medical image analysis tasks is the U-net introduced by Ronneberger *et al.* at MICCAI 2015, which is a CNN with skip connections used to successfully segment microscopy images [125]. The name becomes obvious when one takes a look at its structure seen in Fig. 5.5, the down-sampling and up-sampling operations resemble the letter *U*. The network was designed for segmentation and has quickly become the state of the art for segmentation tasks, but the network architecture is proving very popular for other tasks due to its ability to treat image data on multiple resolution levels. We may note that a parallel can be drawn between the different scales in a U-net and the multi-scale training scheme of VDSR [115], where the U-net architecture removes the need to actually increase the amount of data. This means the great performance of U-nets can be traced back in part to the inadvertent data augmentation done by treating the data on several resolution scales.

CNNs in Super Resolution of Medical Images

The skip connections in the U-net allow high frequency as well as the low frequency information to be processed in parallel and make the architecture suitable for super resolution, for which it has been applied to 3D microscopy in two recent studies. Both the following studies were presented at MICCAI 2017, so we have ranked them arbitrarily. The first study, by Weigert *et al.* [126], uses a U-net to generate a residual image containing the high frequency information which is added to the low resolution input to recover high frequency detail. The network architecture includes fully connected layers, and uses a PSNR loss function. The authors also present a simple architecture following SRCNN [110], which gives worse results. Super resolution of factors $\times 4$ and $\times 8$ are shown to work on synthetic datasets and on real images. The model for real images is trained using a training scheme where

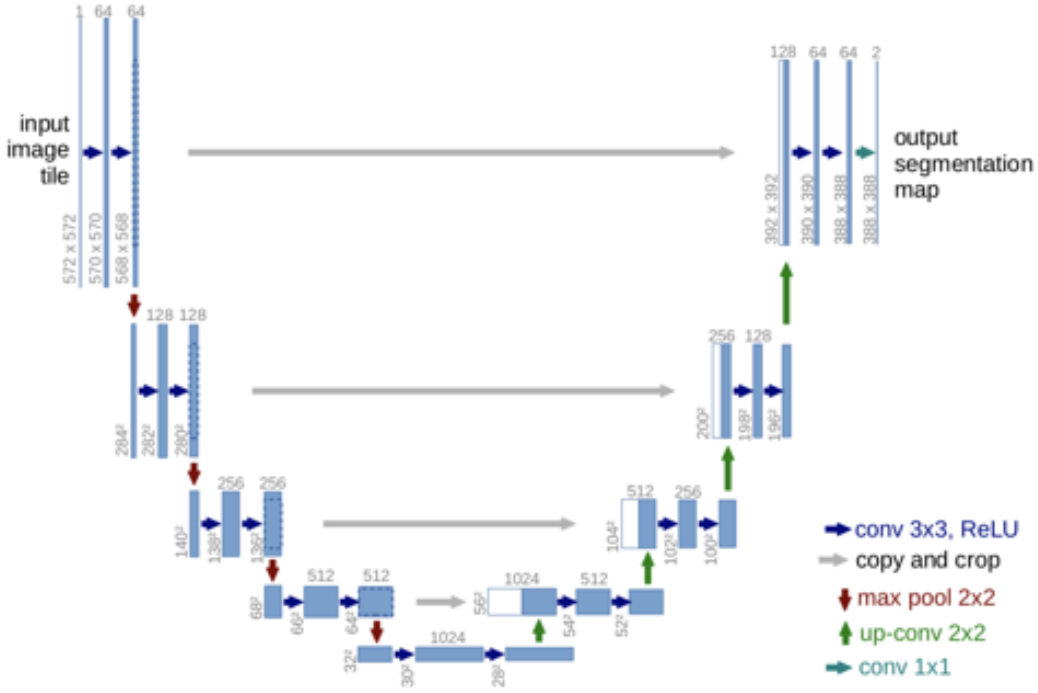


Figure 5.5: The U-net architecture, taken from [125].

the self similarity of the data in different orientations is exploited. Improvements to segmentation performance are shown where the enhanced images approach the segmentation results of the full isotropic ground truth volume. The other study, by Heinrich *et al.* [127], also compares a U-net and a FSRCNN model in 3D used to up-sample synthetically down-sampled anisotropic 3D microscopy volumes. Both architectures are slightly modified to match the demands of 3D data and the larger down-sampling scales, by increasing the depth and sizes of the kernels, and up-sampling is achieved by transposed convolution layers. For the U-net the authors increased the filter numbers and restricted down-sampling to the axis in which the images were isotropically down-sampled. An up-scaling of factor $\times 4$ is achieved. The training data was augmented by separating large isotropic 3D volumes into several pairs of smaller blocks of $64 \times 64 \times 16$ and $64 \times 64 \times 64$ pixel resolutions respectively. Their results show that both architectures can be used for the task with their 3D U-net consistently outperforming their 3D FSRCNN, for example providing better definition of cell membranes [127]. In [128], another 3D network

based on residual learning is presented. This network enhances brain MRI volumes and is trained with data at several enhancement scales, similar to the multi-scale approach of [115].

In [129], the authors describe super resolution of CT images using a residual based CNN and report dramatic improvements to detections of lung modules in the enhanced images. As opposed to using synthetically down-sampled data, this study underlines the importance of real data. In CT this is especially feasible due to fewer motion artifacts. The training is done on sets of images acquired with different slice thicknesses, results of the reconstructions show improvements over interpolations in terms of PSNR and SSIM. Further to that, detection of nodules which only works at the highest of the acquired resolutions works on the enhanced images.

CNNs have has also been applied to super resolution of cardiac MRI by Oktay *et al.* [84], where the authors train two networks: a single image and a multi-image one, to predict residuals which are added to the low resolution image and give its missing high frequency information. Working in a similar manner to FSRCNN and other transposed convolution based approaches, the input to the network consists in solely the stacked short axis images which are up-sampled and super resolved in the network. The authors note they use many stacked $3 \times 3 \times 3$ convolution kernels and ReLU activations, and the first part of the network consists of transposed convolution layers which resamples the input to the size of a high resolution array. For training of the single input network, the authors rely on an impressively large dataset of 3D cardiac volumes. Clinical 2D multi-slice data is used for the training of the multi-input network, after misalignment correction using registration. The authors note slight improvements with the addition of long axis information from few slices in the multi-input network, where masks are used to indicate where data is missing in the long axis views. The modest improvements are noted especially where they overlap with the short axis, which does not come as a surprise. The same group have applied the same methodology to the enhancement of late gadolinium enhanced

cardiac MRI [130]. Both [130] and [84] note superior performance to interpolations and the previously mentioned Multi Atlas PatchMatch (MAPM) [72], which was also developed at Imperial College London. We notice that the results of [130] are performing much better in their quantitative evaluation against interpolations and the MAPM. As a matter of fact this shows a discrepancy between the use of MAPM in both contributions [72] and [130]. In [72], an improvement from 19.9dB to 20.96dB in PSNR is noted between interpolation and MAPM, while in [130] MAPM achieves just a small increase from 36.7dB to 36.6dB from interpolation, and a much higher 40.1dB using the CNN approach. The interpolations appear to be clipped off on the top and the bottom of the images, which may explain this, however the proposed result also suffers from strong artifacts highlighting some shortcomings of that approach. Since the approach works well in [84], there might be an issue with insufficient training data in [130]. More recently, again from the same group of authors, work was extended to an Anatomically Constrained Neural Network (ACNN) that resembles a U-Net and is able to do super resolution and segmentation [88]. The addition of learned shape priors for a global loss added to a pixel-to-pixel loss in the training allows the multi-purpose network to overcome common issues such as misalignment.

Both the progress in research related to CNNs in general and specific to super resolution has evolved in several ways including more complex architectures based on better understanding of functions of each of the elements in networks. Recently, one of the most exciting developments in deep learning has been the development of the previously mentioned GANs, which have enabled to break away from loss functions such as MSE.

5.3 Generative Adversarial Networks

Losses in CNNs, such as the L1 loss, are evaluated on a pixel-to-pixel basis. Such metrics are not necessarily suitable when the input and output image are slightly distorted or misaligned with respect to each other meaning structures of interest do

not occupy the same exact locations in both images. This is a problem consistently occurring in medical imaging, because of patient and breathing motion between acquisitions in cardiac MRI, but also occurs when patients have multiple visits or go through different imaging modalities. Previously mentioned alternative losses included shape regularisation loss [88] and perceptive loss [122]. Research based on the latter lead to impressive advances in deep learning strategies and have culminated in the development of GANs. A GAN, as introduced by Goodfellow *et al.* [131] consists of two networks. The first network is a generator network G, and the second a discriminator network D. The function of G is to generate an output when given a noise vector z as an input, and the function of D is to tell whether the output of G is real, or has been produced by generator G. D outputs a percentage representing the probability that its input, which during training is alternatively the output of G, $G(z)$, is real. Both are trained concurrently (hence the name *adversarial*), where the aim is to minimise the probability of G to generate a wrong label, or $log(1 - D(G(z)))$, and to maximise the probability that D assigns the correct label, or $log(D(G(x)))$. The generator is trained to generate an output which is assessed by the discriminator network. Instead of aiming to minimise a pixel-to-pixel loss, the training process of the generator network aims to minimise the classification error of the discriminator network while the discriminator network is jointly trained to differentiate between real and fake (generated) images. With this training paradigm, the generator is trained to produce images that can fool the discriminator at the same time as the discriminator becomes better at differentiating between real and fake outputs. The first GANs were used to generate numbers from random noise vectors. A few months after [131] was first online, conditional GANs were introduced in [132]. Conditional GANs build on GANs with the addition of external information y, where G and D are conditioned with additional information. This is done by inputting the information to both G and D, a conditional probability $log(1 - D(G(z|y)))$ and $log(D(G(x|y)))$ results from this addition. Different architectures for discriminators have been put forward, such as the PatchGAN in [133], where features of patches are assessed instead

of the image as a whole to make a decision on whether the generated output is real or fake. It has been shown that PatchGANs require fewer parameters and lower computing power. An additional condition may be set upon a conditional GAN, where the generated outputs have to match the inputs. These conditional GANs learn a structured loss which penalise the output based on the input, this can be a L1 loss. The result of this is that G not only has to deceive D but also be close to the input image, and the difference is that the input is not a noise vector.

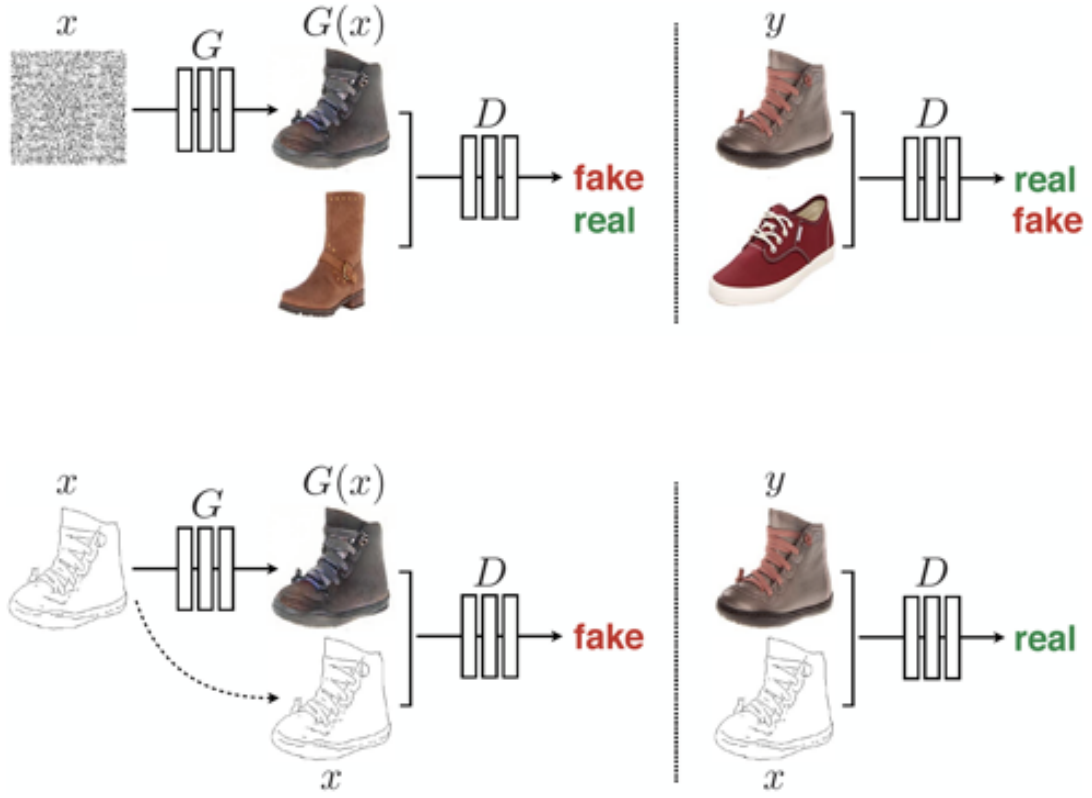


Figure 5.6: Top: a Generative Adversarial Network. Bottom: A conditional Generative Adversarial Network for image-to-image translation, both adapted from [134].

A combination of conditional GANs, the U-net [125], and the PatchGAN discriminator have led to the development of a novel image-to-image transfer technique at the core of the pix2pix software and is presented in [134]. Fig. 5.6 shows the

concept of GANs and conditional GANs. In the former, a random noise vector is used as input to the generator whereas in the latter an edge image is used as input to the generator as well as to the discriminator. A U-net architecture is used for the generator and a PatchGAN classifier is used for the discriminator. This particular conditional GAN learns a structured loss which penalises the output based on the input image, and it transforms input images into output images of the same scene but with different appearance or texture. Examples show the ability of going both ways from edge images to real images including shoes and handbags, from labeled segmentations to real pictures, from satellite images to drawn maps, as well as colorising images and changing the daytime appearance in photos. In the the conditional GAN described in [134], the overall objective is given by

$$G* = \arg\min_{G}\max_{D} \mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_1(G) \tag{5.1}$$

where

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[logD(x,y)] + \mathbb{E}_{x,z}[log(1-D(x,G(x,z))] \tag{5.2}$$

Generative Adversarial Networks in Super Resolution

With a quickly moving field, GANs have reached super resolution approaches soon after their initial publication. Photo-realistic natural image super resolution is achieved in the previously mentioned [119] using SRResNet as the generator. The aim of the authors was to recover finer texture details for large scale up-sampling. Their SRGAN achieves super resolution of factor $\times 4$. As previously mentioned, the visual quality is not matched by quantitative evaluation, as a matter of fact the authors show that PSNR and SSIM for interpolations are higher than that of SRGAN. As with [122], which was also using perceptual loss but not with GANs, the quantitative evaluation does not match the visual superior quality of the results. This correlates with the fact that the smoothness promoted by L1 does not necessarily match enhancements to perceptive quality. An enhanced SRGAN (ESRGAN) has been proposed in [135], the authors note that the enhancement

resulted from their removal of batch normalisation, as has been done for EDSR [120], as well as additional residual connections within residual blocks. Qualitative results are, subjectively, enhanced but the numerical evaluation again does not reflect any consistent improvement against SRGAN or interpolation.

Generative Adversarial Networks in Medical Image Analysis

GANs have found application in medical imaging for example to enable image modality transfer from MRI to computed tomography using unpaired data [136]. GANs have also quickly become popular for data augmentation tasks [137, 138]. The applications of GANs to compressed sensing may allow the acceleration of MRI acquisitions such as in the following [139]. The authors feed real and imaginary parts of the under-sampled brain MRI k-space data into the GAN which generates the missing complimentary k-space data, before the discriminator decides on the image quality after transforming the k-space into image space. Both networks are sequential with 6 blocks of convolutions, batch normalisations and ReLU for the generator, and 4 for the discriminator with the addition of max pooling layers and a fully connected layer for the classification. Their results show enhancements in PSNR as well as qualitative aspects. On the other end of the pipeline, super resolution of brain MRI using GANs is explored in [140]. The architecture of the latter is described as a network comprised of multiple blocks with additional skip connections than normal residual blocks, similar to [120]. Low resolution data is generated by truncating the k-space by a factor of $\times 2$ before being interpolated to match the size of the high resolution image. In [141], a two-stage procedure is followed. First, a reconstruction network, which is a sequential CNN, is trained with MSE loss. Secondly, the authors introduce a refinement network, a residual U-net, and a VGGNet based discriminator. The training and testing data consists of late gadolinium volumes and 2D data, it is not explicitly stated whether or not the 2D images are synthetic. Results on images subject to an 8-fold up-sampling of k-space show improvements in PSNR as well as segmentation scores. A more

in depth review of the use of GANs in medical image analysis may be found in the following references: [142, 143].

5.4 Recurrence in Deep Learning

While there are many more topics on deep learning that we could be reviewing, the last section of deep learning we cover here is on the concept of recurrence. We chose this because of the dynamic nature of cardiac cine MRI, which acquires images in time sequences, and by the use of this concept in our own work as presented in chapter 7.

5.4.1 Recurrent Neural Networks

In several different kinds of sequential data, including natural speech or videos, a concept of time or continuity within the data is involved. The CNNs mentioned before work solely on static information, for example a single image. CNNs can be applied to every individual element in a sequence, but would not be able to use previous information to infer anything on upcoming information in a such sequence. Volumetric data including dynamic images (2D+t) such as cardiac cine MRI sequences, or 3D images, contains strong spatial dependencies between its adjacent slices and makes it a good fit for the use of such networks.

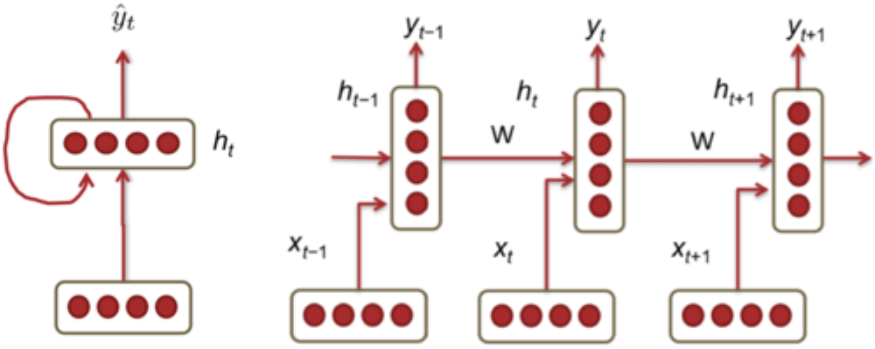


Figure 5.7: A RNN loop on the left and in its unfolded form on the right, adapted from [144]. A box in the figure on the right hand side is a hidden layer at a given time step, t.

A recurrent Neural Network (RNN) can be used to maintain past information

by acting as a loop. RNNs can be thought of as repeating copies of the same network, which passes the data to an adjacent network. Fig 5.7 shows a RNN loop and an unfolded representation of that loop. A RNN is described by the following equations [144]:

$$h_t = W^{(hh)} f(h_{(t-1)}) + W^{(hx)} x_t \tag{5.3}$$

$$\hat{y} = f(h_t) \tag{5.4}$$

Where x_t is a sequential signal with t steps, h_t the hidden layer output features, $W^{(hh)}$ are the weights governing the contribution of the previous time step h_{t-1}, $W^{(hx)}$ the input weights, $W^{(s)}$, and f an activation function giving the output $\hat{y}$ of the layer, often these are sigmoids or the hyperbolic tangent function. The derivatives of the very similar looking functions (Fig 5.2) are both quite similar and resemble short impulses and are otherwise equal to 0.

5.4.2 Long Short-Term Memory

It has been noted that when using many time steps, computing the gradients backwards many times through those functions that have gradients that are mostly 0, results in the so-called vanishing gradient problem and makes the networks difficult to train. To address this issue of long term dependence, Long Short Term Memory (LSTM) networks have been introduced in [145]. Like many concepts in deep learning, LSTM networks are actually not that recent and date from the 1990s. LSTM networks address the vanishing gradient issue by the addition of a *memory unit* c and a set of *gates*: i (input), f (forget), and o (output). A somewhat similar concepts are Gated Recurring Units (GRU) [144]. In [146], Convolutional LSTM networks with the ability to handle data in arrays are introduced, and used for short term meteorological forecasting. Convolutional LSTMs are governed by the following equations:

$$i^{(t)} = \sigma(W^{(xi)} * x^{(t)} + W^{(hi)} * h^{(t-1)} + W^{(ci)} \circ c^{(t-1)}) \tag{5.5}$$

$$f^{(t)} = \sigma(W^{(xf)} * x^{(t)} + W^{(hf)} * h^{(t-1)} + W^{(cf)} \circ c^{(t-1)}) \tag{5.6}$$

$$c^{(t)} = f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tanh(W^{(xc)} * x^{(t)} + W^{(hc)} * h^{(t-1)}) \tag{5.7}$$

$$o^{(t)} = \sigma(W^{(xo)} * x^{(t)} + W^{(ho)} * h^{(t-1)} + W^{(co)} \circ c^{(t)}) \tag{5.8}$$

$$h^{(t)} = o^{(t)} \circ \sigma(c^{(t)}) \tag{5.9}$$

Where $*$ is the convolution operation, $\circ$ denotes an element wise product, σ an activation function. h is the output, o its activation, and W are the weights which are learnt at the training stage. Larger kernels may be used to capture faster motions whereas smaller kernels may be used to capture slower motions.

5.4.3 Applications of Recurrence in Deep Learning to Medical Image Analysis

RNNs and LSTM networks are starting to find applications in medical image analysis. Other than in a temporal sense, recurrence can also be applied in a spatial sense, by considering adjacent slices in a 3D image. In a study on prostate MRI for cancer segmentation, adjacent 2D slices were fed into a U-net fitted with recurrent layers at every convolution [147]. RNNs have first been applied for the processing of cardiac cine MRI by Poudel *et al.* [148], where RNN blocks were included at the lowest resolution level of a U-net to take advantage of low frequency features in consecutive frames of the cardiac cycle. RNNs have also been applied to compressed sensing of dynamic MRI in [149]. LSTMs have been applied to enhance performance of myocardium segmentations in cardiac cine MRI sequences [150]. In that study, similar to [148], recurrent layers are present in the lower resolution levels of the network architecture. In [151], LSTMs are used for frame identification in a cardiac MRI sequence.

In this chapter we have presented the reader with background on deep learning and CNNs, with a focus on general super resolution, the super resolution of medical images, as well as generative models and recurrence in deep learning. The next chapter will deal with super resolution reconstruction of long axis slices from short axis data using deep learning methods.

6

Convolutional Neural Networks for Long Axis Slice Reconstruction from Short Axis Stacks

In the previous chapter, key concepts in deep learning including CNNs, GANs and LSTMs, as well as a literature review on their applications to super resolution and medical image processing were presented to the reader. This chapter presents applications of deep learning approaches to static frames of 2D multi-slice cardiac cine MRI data. We start by applying some benchmark CNN architectures and compare their performance on synthetic low resolution data from long axis slices, before moving to the use of GANs in order to combine actual short axis acquisitions with corresponding long axis data, and finally concluding on transfer learning using data from a single heart to reconstruct additional arbitrary long axis slices which are not commonly acquired. Some of the work in this chapter has been presented at ISBI 2018 in Washington DC, USA.

6.1 Introduction

A recent deep learning based effort has given automatic measures of cardiac function [7, 8], and a recent challenge at MICCAI 2018 was dedicated to this task[1]. Efforts like this may open the possibility of fully automated interpretation methods in the future, which would take the images as an input and produce simple metrics such as ejection fraction. However, the interpretability of CNNs for medical image analysis is important, as in clinical settings such "blackboxes" are not commonly accepted. Instead, we believe it is important to show the steps that have brought a specific result. This thesis aims at contributing to this by producing intermediate results (new slices) that facilitate the analysis task.

Interestingly enough, a number of top performing methods proposed to the aforementioned challenge do just that: instead of directly estimating clinical metrics, they first segment the images and then proceed to compute the metrics from the segmentations [2]. We allow ourselves to draw a parallel to the initial frustration of the scientific community with the efficacy of CNN "blackboxes". The excellent results achieved by CNNs made them widely embraced and they have been making up the a major part of research in computer vision ever since. Once their inner workings and meanings of the hidden layers became more intelligible, CNNs became even more mainstream and results improved significantly. Similarly, we believe that clinical data can be used in conjunction with deep learning to produce state of the art results for anatomical and functional assessment out of the box, but it can also be divided into several steps. From a low resolution image, a higher resolution image can be reconstructed from which a more accurate segmentation can be obtained by a clinician or a subsequent network. Clinicians can make a decision if they are convinced by the individual steps or not. In addition to

[1] The website for the challenge may be visited here `https://lvquan18.github.io`. The proceedings are not yet available.

[2] Our group has participated in the challenge which is how we can say this while the proceedings are not yet available, and can therefore not reference this yet.

that, using the sequential data in its context can help determining metrics such as ejection fraction or enable motion analysis.

6.2 Data

In our work described in this chapter, as well as in chapter 7, we used the freely available data from the Second Annual Kaggle Data Science Bowl. This is, at the time of writing this thesis, the largest challenge in terms of the amount of publicly available 2D multi-slice cardiac MRI sequence data. The problem participants were asked to solve for the challenge was to automatically determine the ejection fraction given only multi-slice cardiac MRI sequences and ground truth ejection fraction values for the training dataset. All of the top ranking teams used deep learning approaches, including U-nets for segmentation [152]. The dataset covers acquisitions of over five hundred hearts. Every data set has 30 frames, however the number of short axis and long axis slices differ, and the data was acquired using different scanners at several centres. Vertical and horizontal long axis slices were present for most of the patients, however a non-negligible part of the data had only one, or no long axis views. The patients have a large spread of age and heart sizes, which is advantageous to preserve the generalisation properties of our experiments. We discarded unusable data (e.g. that was affected by very strong artefacts, wrongly labeled as being long axis slices, or data without long axis slices), by visual inspection. The discarded data totalled to 7.8% of the entire available images. We performed experiments with two different sets of training data. For the first one, we used the long axis slices to generate synthetic low resolution images, resembling cuts through a stack of thick slices. This allowed us to use the original high resolution slices as ground truth for quantitative validation. In the second experiment, we used short axis data as the low resolution input, and reconstructed high resolution long axis slices. In this way we incorporated all aspects of real acquisition, with the downside that no direct comparison with ground truth was possible due to the slice-to-slice misalignments and variations in image contrast. The processing for both is explained in detail below. All of the images were resampled

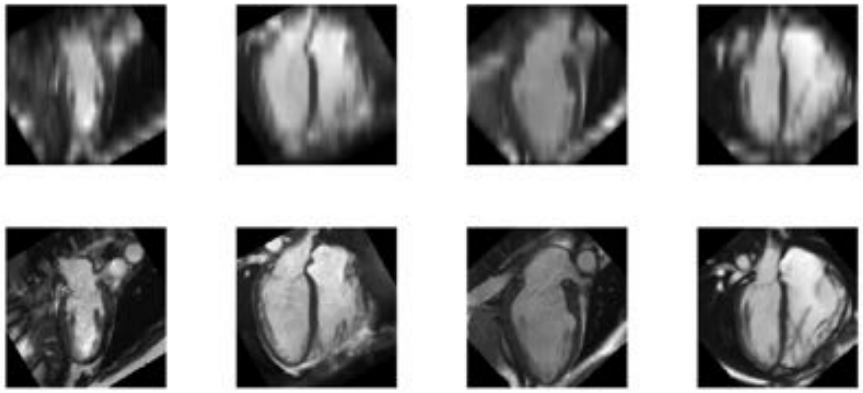


Figure 6.1: The top row shows 4 examples of synthetic low resolution long axis data, showing two horizontal and two vertical long axis slices. The bottom row shows the corresponding ground truth acquisitions.

to isotropic resolution of $1.4mm \times 1.4mm$, and we cropped the images to 128×128 pixels around the approximate centre of the left ventricle, which was determined by manual inspection. Every image was also normalised such that all pixel values lie in the intensity range between 0 and 1.

6.2.1 Synthetically generated low resolution data

For the synthetic data, we used the long axis slices exclusively. We ensured all the images have the same upright orientation where the base of the heart is located at the top of the image and its apex points towards the bottom by applying rotations to the images. The angles of rotation required for this alignment were obtained from the metadata of the DICOM files. We down-sampled every image in the baso-apical direction to match the slice thickness of standard short axis slices of 10mm. In this way, we reconstructed images similar to those that would be given from the short axis stack. The final step was to apply bilinear (since the long axis slices are 2D) interpolation to the data. We show an example of synthetic training data in Fig. 6.1. Combining all the available vertical and horizontal long axis slices, where a total of 1114 sets were available with 30 frames per sequence, resulted

in 33420 low-high resolution image pairs. We augmented the data by a factor of $\times 2$ by applying a horizontal flip resulting in a total of 66840 pairs. We decided against other transformations because we wanted to preserve the orientation of the ventricles with the apex towards the bottom and in axis vertically with the base. We did not perform any rotation or scaling to further augment the data. The way the interpolated short axis low resolution images relate to the high resolution long axis acquisition is by an increase of resolution in the slice selection direction only, therefore we ensured the data maintains this relation. Our aim is that the network is able to reconstruct isotropic long axis slices from consistently anisotropic short axis in parallel. We believe that since the super resolution is just in the through plane direction, it makes sense to keep the data consistently anisotropic in the same orientation and scale. We also did not differentiate between horizontal and vertical long axis views, we included both together in the training, validation and testing datasets. In this way, we aim to demonstrate the ability of the method to recover images with different appearance (e.g. in terms of the number of chambers), with the eventual goal of using the network to reconstruct slices in any arbitrary orthogonal orientation from short axis stacks.

6.2.2 Combination of short and long axis slices

For the second set of experiments, which used the short axis stack low resolution input, we applied rotations to every heart such that images from the short axis stack were in plane of a 3D volume array. We show this in Fig. 6.2, where the left hand side shows the real world coordinate system and the right hand side shows a coordinate system with the short axis parallel to the x-y of a 3D volume array. We did this to be able to systematically re-sample the short axis data of all hearts into a 3D array in the same orientation. A 3D volume with isotropic resolution was generated by interpolating the short axis images using linear interpolation. For every available long axis slice, a 2D synthetic image was extracted from the interpolated short axis volume at the location of the long axis slice. This was done using the acquisition model used for the method presented in chapter 4, using the meta-information of

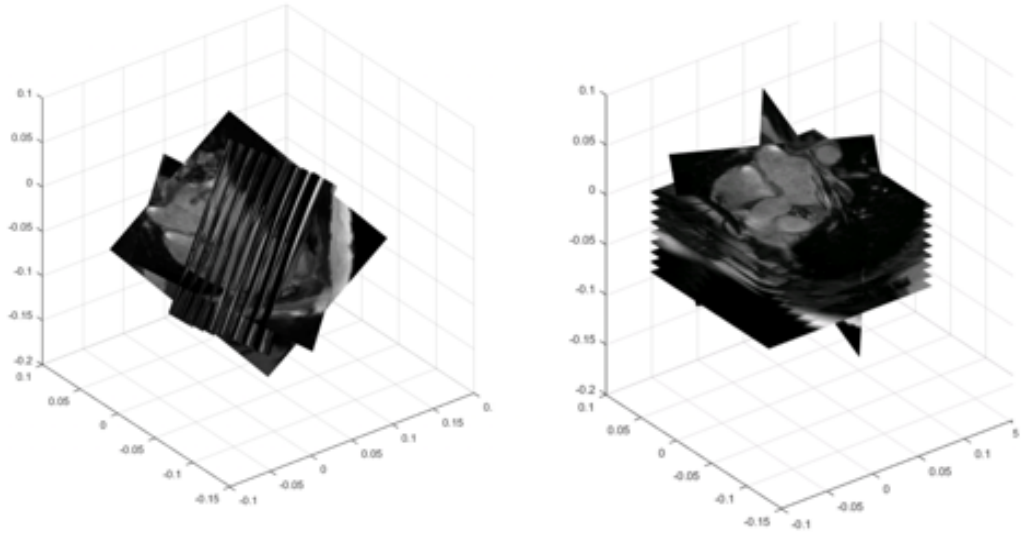


Figure 6.2: Left: DICOM data in its original orientation. Right: the same images after applying rotation.

the slices to get the positional information. Just as for the synthetic data, 66840 pairs of low-high resolution images were available after augmentation by horizontal flipping, and including all frames for every case. As a matter of comparison, we have more data available for our experiments than the amount used in landmark contributions such as SRCNN (24,800 after sub-images were generated) in [110] or the U-net (30 512×512 images, augmented by a factor of 7) in [125]. Fig. 6.3 shows the combination of short and long axis data. In the long axis slice reconstructed from the short axis stack, the slight misalignment highlighted by the red arrows is noticeable in the ventricular septum, the absence of papillary muscles is visible and the apex is not very well defined as is the myocardium at the bottom right of the left ventricle (which is on the right hand side, and the larger of the two ventricles).

6.3 Methods

We present the methodology involving the synthetic and short axis interpolated data on established architectures which we have slightly altered. First the SRCNN, that

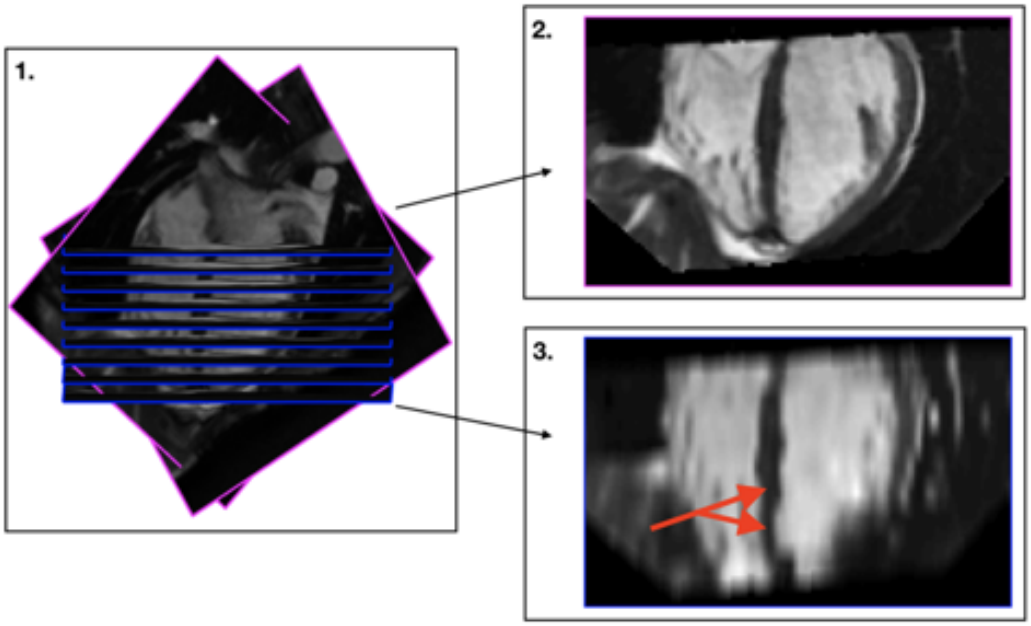


Figure 6.3: Extracting a long axis slice through the short axis stack. 1. View of short axis stack (blue) and long axis slices (pink) in 3D space. 2. High resolution long axis 4-chamber view. 3. View of the *approximately* equivalent position through the short axis stack. Red arrows are pointing to slice misalignment.

has been designed for super resolution, secondly the U-net that has been designed for segmentation but has been successfully implemented for various other tasks, including super resolution. We finally present the use of conditional GANs to super resolution of cardiac MRI, where we have also slightly altered the architecture to work with our data. We present the details below.

All of the networks were implemented in Python using the Keras library [153] running with Tensorflow backend, and training was performed on a Nvidia GeForce GTX 1080 Ti 256 RAM GPU. For optimisation, we always used the Adam algorithm with the recommended values for $\beta_1 = 0.9$, $\beta_1 = 0.999$, and $\epsilon = 10^{-8}$ [96]. We use the synthetic data and the interpolated short axis data, both sets consisting

of the low resolution and high resolution image pairs as respective input and labels, as presented above.

6.3.1 SRCNN

After a brief review of the literature at the time this task was started (in mid 2017), we decided to experiment with deep learning methods and so it made sense to start with the seminal SRCNN [110], especially given our initial project with dictionary learning, which we have described in Section 4.1. SRCNN deals with isotropic super resolution of interpolated 33×33 pixel natural images by a factor of 3 and 4 in both directions. We deal with 128×128 pixel images and an anisotropic super resolution of factor $\times 8$ in one direction, while the other remains the same. Larger kernel sizes result in larger receptive field, meaning the information used for the super resolution of each pixel increases. After the first attempt where we used the original kernel sizes, we chose to use larger kernels as it was done in [127], since we are looking for a larger up-scaling factor and use larger images than the original study. Instead of the $5-1-9$ kernel sizes we tried several incrementally larger sizes and ended up using $9-1-17$, which is about double of the original sizes, and doubled the numbers of filters to $n_1 = 128$ and $n_2 = 64$.

6.3.2 U-net

After having gained some practical experience with CNNs via implementing SRCNN, we started considering different more complex architectures. The U-net and its ability to treat low and high frequency features in parallel appeared to be very interesting in super resolution applications, where the aim is to recover high frequency information when using an input containing the low frequency features. In this experiment, we used the following architecture: $C_{64} - M - C_{128} - M - C_{256} - M - C_{512} - M - CD_{1024} - CD_{512} - U - CD_{256} - U - C_{128} - U - C_{64}$. Here C_x represents a block consisting in two successive 3×3 convolutions with batch normalisation and leaky ReLU activations, the subindex x the number of filters, M a 2×2 maximum pooling operator, and U a 2×2 up-sampling. The layers with the same filter

numbers are connected by skip connections. The architecture also incorporates dropout layers D. Here the dropout rate is 50%, which indicates the probability that a neuron is randomly switched off for a certain weight update. This is the same architecture used for the U-net used in [134], but with 3×3 convolution kernels and stride 1, and with smaller input size and fewer channels, as the data we use are grayscale images and smaller in size.

In addition to that, we have also implemented the same architecture with convolutions of kernel size 4×4 and stride 2, which is the architecture of the generator network in [134]. The paper however lacks an explanation as to why kernel sizes of 4×4 and stride 2 are used, which seems unusual, and worth mentioning given the rare use of even convolution kernel sizes. An explanation is given in [154], which uses conditional GANs for face generation. The point the authors make, is that if the kernel size of the transposed convolution is not a multiple (different than 1) of its stride value, the output regions resulting from those transposed convolutions will generate uneven overlap which can result in artifacts in the final outputs. An illustration of the issue is shown in Fig. 6.4, where the overlaps created by a 3×3 kernel with a stride 2 transposed convolution are displayed. Here, every pixel affects a small 3×3 region represented by the light blue. The tiling with this filter size and stride results in uneven distribution of overlap, shown by different shades of grey. We note that we have not experienced this issue in any of our reconstructions.

We train the following models: $U - net_{LA3}$ and $U - net_{LA4}$ on synthetic data where the former has filter size 3 and the latter filter size 4, and $U - net_{SA3}$ and $U - net_{SA4}$ on interpolated short axis data with long axis ground truth. This gives us a total of four U-net based models. We use Adam optimisation with a learning rate of 4.5×10^{-4}, for 50 epochs. The learning rate was evaluated empirically aided by the results obtained on the validation data. This learning rate gave the best results and fastest convergence whilst doing a sweep of various learning rates between 10^{-1} and 10^{-6}.

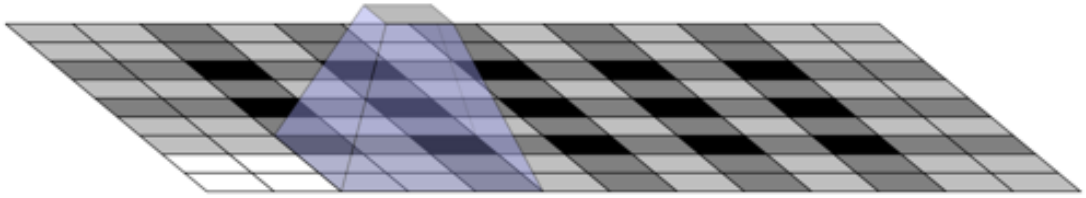


Figure 6.4: Overlap resulting from transposed convolution. Taken from an illustrative animation on https://distill.pub/2016/deconv-checkerboard/.

6.3.3 Conditional GAN

In this section we present long axis slice reconstruction from interpolated short axis data, one of the main contributions of this chapter and thesis. After we obtained reasonable results with the U-net and synthetic data, we wanted results reconstructed from the acquired short axis data, and to reconstruct images that are not overly smooth in appearance. Motivated by GANs and their ability to go beyond pixel-to-pixel loss, we decided to implement that idea to overcome the misalignment between short and long axis slices. The U-net used as a generator is the same as described above in subsection 6.3.2, with convolutions of kernel sizes 4×4 and stride 2. For the discriminator, we have altered the architecture from the one used in [134] to change the receptive field size of the network. The authors of [134] note that when the patches resulting from the receptive field size are too small, artifacts are present in the result, and also that when the patch size was too large there was no improvement over just using the entire image. It is also noted that the smaller patches make the network less computationally expensive. In their implementation, the convolutions resulted in a receptive field size of 70×70 for an input size of 286×286 pixels. By changing the strides in the last convolutional layer from 2 to 1 we reduced the receptive field size to 37×37 pixels. We have done

this because the images we are using are about half in number of pixels per side compared to the data used in [134], 128 against 286 pixels. The final discriminator architecture we used was as follows: $C_{64,4,1} - C_{128,4,2} - C_{256,4,2} - C_{512,4,1} - C_{1,4,1}$, where $C_{n,k,s}$ is a convolution with a kernel of size $k \times k$, n filters and a stride of s, followed by a leaky ReLU. The final activation is a sigmoid function, leading the discriminator to give a value between 0 and 1, representing the probability of the input being real or generated. We have trained two conditional GANs: one with the synthetic data generated from long axis slice alone (LAGAN), to prove the concept that a GAN can generate images with a more realistic texture, and not overly smoothed in appearance, like the results obtained from the non-adversarial CNN such as the $U - net_{LA3}$ model. We train another GAN using the long axis slices extracted from the interpolated short axis data with the aim of reconstruct missing or additional long axis slices. We refer to this model, which incorporates short and long axis slice information, as SALAGAN. The value of λ in equation (5.1) was set to 100, the learning rate to 2×10^{-4} and the momentum parameters to $\beta_1 = 0.5$ and $\beta_2 = 0.999$, these are the same values as used in [134]. Both the U-nets, with both kernel sizes, have fairly large parameter numbers originating from the high filter numbers. Both U-net models have around $\mathcal{O}(10^7)$, where the model with larger convolutional kernels (4×4) carries slightly more parameters.

6.3.4 Transfer learning

Fine tuning networks to become data specific is a technique commonly used when large labeled training sets are not available, as the first layers in deep neural networks are often not specific to certain tasks or data. Often it is only the last layers that display the task specific features, as shown by a study on transferability of features in deep neural networks [155]. It is possible to fine-tune networks by retraining the network on a different data set, this can be done by either allowing all the layers to be retrained, or by keeping specific layers as they are (also called freezing layers). Often the first layers are kept frozen. If the target data is close to the data used to train the original network, time can be saved by keeping the first layers frozen. In

recent work by Bai *et al.* [156], clinical metrics for cardiac MRI are automatically evaluated by a fully convolutional network. The study is based on UK biobank data, and the authors point out that the network fails to produce adequate results on different data sets. Indeed, the UK biobank data is relatively homogeneous as it consists of a majority of healthy individuals of a similar age group, strict imaging protocols and relying on the same scanner models [35]. which prompted them to fine-tune their network on other data, which resulted in improvements of the computed metrics. In the paper, it is not specified whether certain weights were frozen or not.

For this last experiment on deep learning approaches using static frames we use the conditional GAN we have trained on short axis data (SALAGAN) as a pre-trained network for transfer learning, to fine tune the generic network and make it data specific to one particular heart or patient. We take all the frames from two long axis slices from one heart and extract their equivalent slices through the short axis stack. With augmentation, this gives us 120 high-low resolution image pairs. For every one long axis training view, we have 30 frames, and a final number of 60 samples for both the long axis acquisitions, by applying a horizontal flip. We train the model, called T-GAN, with weights initialised from the SALAGAN model on this data, with the aim to be able to reconstruct an additional third long axis slice given the short axis stack and the two long axis slices of a particular patient. Here the aim is not that the network may be perfectly able to reconstruct the long axis slices that are already present, but by learning the mapping between short and long axis slices of one or two views of the same heart, it would aid the reconstruction of arbitrary other views orthogonal to the interpolated short axis stack. We believe the newly reconstructed images would have to be orthogonal to the short axis stack, because all of our training data consists in views from orthogonal cuts through parallel short axis stacks.

With the aim of clarity, we assign names to all the models described above. A summary of the names of the networks can be seen in Table 6.1. The models dealing with synthetic data only are referred to as "LA" because they are only using the long axis data as output, and the synthetic low resolution version as input.

Underlying architecture	SRCNN	U-net	U-net	conditional GAN	(Transfer) conditional GAN
Filters sizes	9-1-17	3×3	4×4	4×4	4×4
LA input and output	$SRCNN_{LA}$	$U-net_{LA3}$	$U-net_{LA4}$	LAGAN	
SA input, LA output	$SRCNN_{SA}$	$U-net_{SA3}$	$U-net_{SA4}$	SALAGAN	T-GAN

Table 6.1: Summary of all the networks we have trained: underlying architectures, filter sizes, data used, and names for each model.

6.4 Results

6.4.1 Results on synthetic down-sampled long axis data

Table 6.2 shows quantitative results via PSNR and SSIM values for interpolations against the networks trained on synthetic data including SRCNN, the two U-nets with different kernel sizes as well as the conditional GAN.

	Interpolated	$SRCNN_{LA}$	$U-net_{LA3}$	$U-net_{LA4}$	LAGAN
PSNR	22.21 ± 0.66dB	23.91 ± 0.55dB	26.10 ± 0.92dB	25.95 ± 0.54dB	25.38 ± 1.40dB
SSIM	0.725 ±0.032	0.769 ± 0.028	0.799 ± 0.027	0.799 ± 0.026	0.804 ± 0.017

Table 6.2: Quantitative evaluation of the reconstructions given by several models trained on synthetic data. PSNR, SSIM for interpolations, $SRCNN_{LA}$ and two synthetic U-net models.

With Fig. 6.5, we show results using long axis examples for the networks and interpolations: $SRCNN_{LA}$, the two U-nets ($U-net_{LA3}$ and $U-net_{LA4}$), and $LAGAN$, all trained on the synthetic data obtained from anisotropic down-sampling long axis slices.

Yet another comparison showing interpolations and results from networks trained on the synthetically generated low resolution data is shown for one 2-chamber unique view in Fig. 6.6. We show the interpolation in the top centre image and a model $U-net_{LA3}$ result on the bottom centre image. Here the apex is not too severely affected by the equivalent short axis stack sampling, however the $U-net_{LA3}$ gives result superior when visually compared to $SRCNN_{LA}$ and the interpolations. The result given by LAGAN resembles the appearance of the ground truth more closely, yet areas outside the heart cavity could be prone to artifacts.

6.4.2 Results on interpolated short axis data

Means and standard deviations of PSNR and SSIM are shown in Table 6.3, to evaluate our results quantitatively. It compares interpolations against the networks

	Interpolated	$SRCNN_{SA}$	$U-net_{LA3}$	$U-net_{LA4}$	SALAGAN
PSNR	21.30 ± 0.66dB	20.33 ± 1.33dB	19.82 ± 0.72dB	19.55 ± 0.68dB	22.39 ± 2.55dB
SSIM	0.690 ± 0.614	0.616 ± 0.037	0.620 ± 0.031	0.620 ± 0.027	0.786 ± 0.028

Table 6.3: Quantitative evaluation of the reconstructions given by several models trained using short axis input data input, and ground truth long axis data output. PSNR, SSIM for interpolations, $SRCNN_{SA}$, $U-net_{LA3}$, $U-net_{LA4}$, and the conditional GAN models.

trained on interpolated short axis data, including SRCNN, the two U-nets with different kernel sizes as well as the conditional GAN.



Figure 6.5: Interpolations and results for four long axis views, using $SRCNN_{LA}$, the two synthetic U-net models, $U-net_{LA3}$ and $U-net_{LA4}$, $LAGAN$, as well as the ground truth. The figure is flipped to landscape orientation for larger display.

Fig. 6.7 shows the result of a long axis slice reconstruction from short axis data using interpolation, $SRCNN_{SA}$ and from using $U-net_{SA4}$. In Fig. 6.8 and 6.9, we show results of the conditional GAN on real short axis data. The former includes a two and a 4-chamber view scan where the apex is not well defined through the short axis, while the latter shows a more extreme case for a 4-chamber view.

Since the ground truth data is not available for arbitrary additionally reconstructed long axis slices using the T-GAN model, we show the data consistency of the proposed results by drawing intensity profiles of the overlap between the planes occurring between the acquired and reconstructed long axis slices. We show two

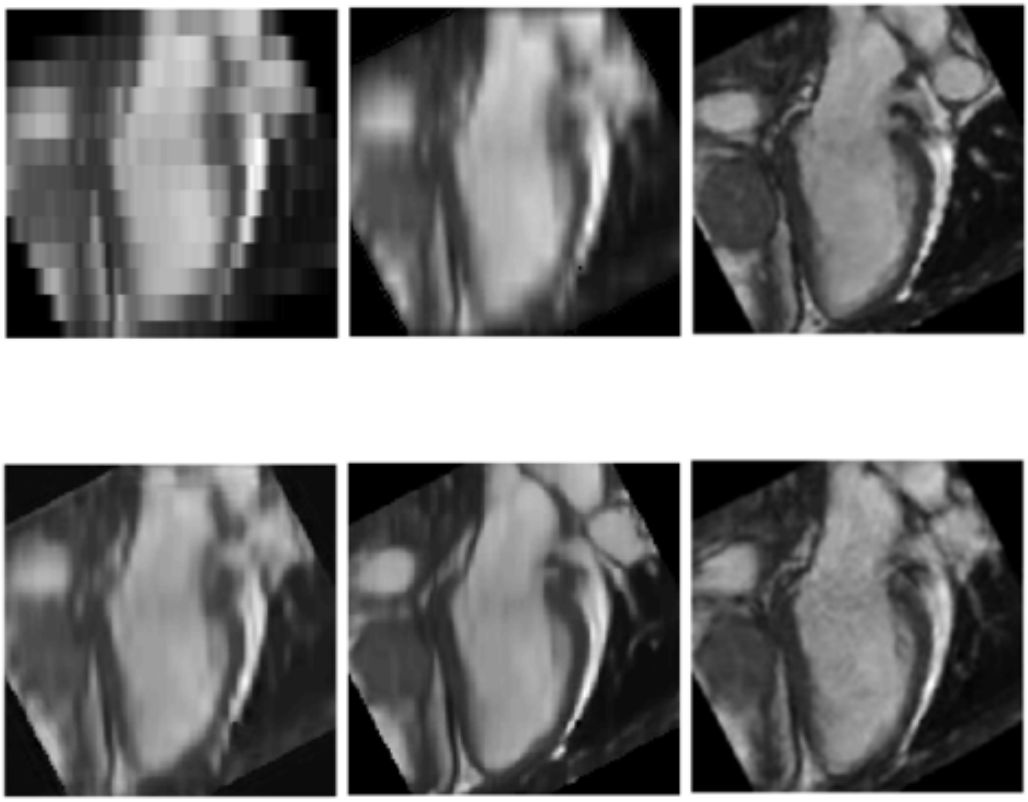


Figure 6.6: Results on synthetic data of a 2-chamber long axis view. Top row from left to right: nearest neighbour interpolation, bilinear interpolation, ground truth. Bottom row from left to right: $SRCNN_{LA}$, $U-net_{LA3}$, LAGAN.

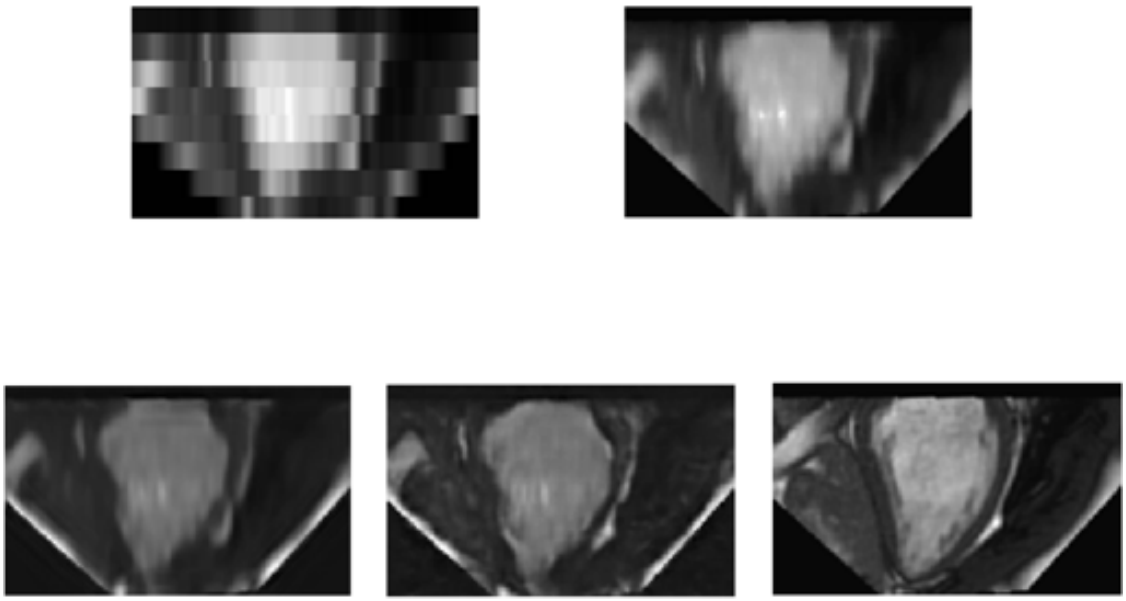


Figure 6.7: Interpolation and results for one long axis, using short axis interpolation in training, for SRCNN as well as a U-net model. Top left: nearest neighbour interpolation, top right: linear interpolation, bottom left: $SRCNN_{SA}$, bottom middle: $U - net_{SA4}$, bottom right: ground truth.

views of the resulting long axis slice with the original slices in Fig. 6.10. This is where we show that the method can reconstruct slices that are not typically acquired in clinical datasets. The profiles shown in Fig. 6.11 show the intersections between the additional slice and the 2-chamber and 4-chamber views seen in Fig 6.10. The intersections are also highlighted in Fig. 6.10 where the orange and blue lines represent the lines along which the profiles were taken. Features such as the myocardium represented by dips in intensity are highlighted by arrows in Fig. 6.11.

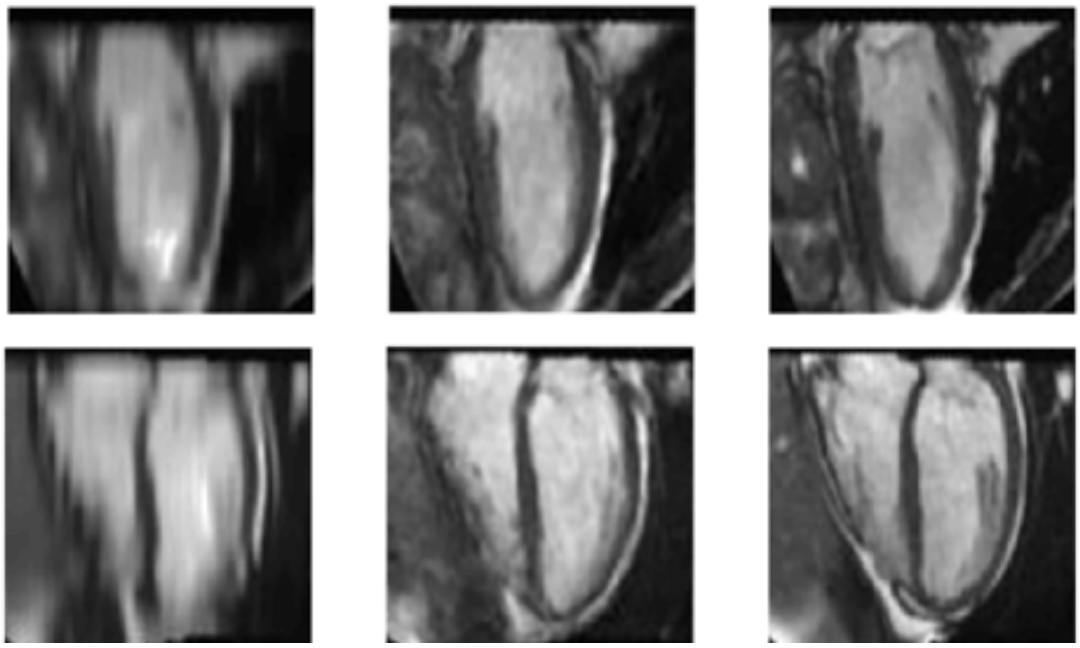


Figure 6.8: Result using real short axis data as input, for SALAGAN. From left to right: short axis input, result using SALAGAN, ground truth. Top: 2-chamber view, bottom: 4-chamber view.

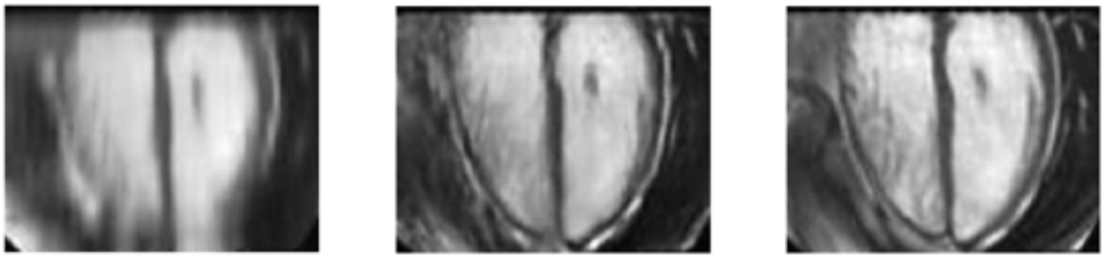


Figure 6.9: Result using real short axis data as input, for SALAGAN. From left to right: short axis input, result using SALAGAN, and ground truth.

6.5 Discussion and Conclusions

The results are summarised in the above section. We can see the resulting images as the most statistically probable long axis slices, given a particular short axis stack, based on the statistics drawn from a large population of examples. Different

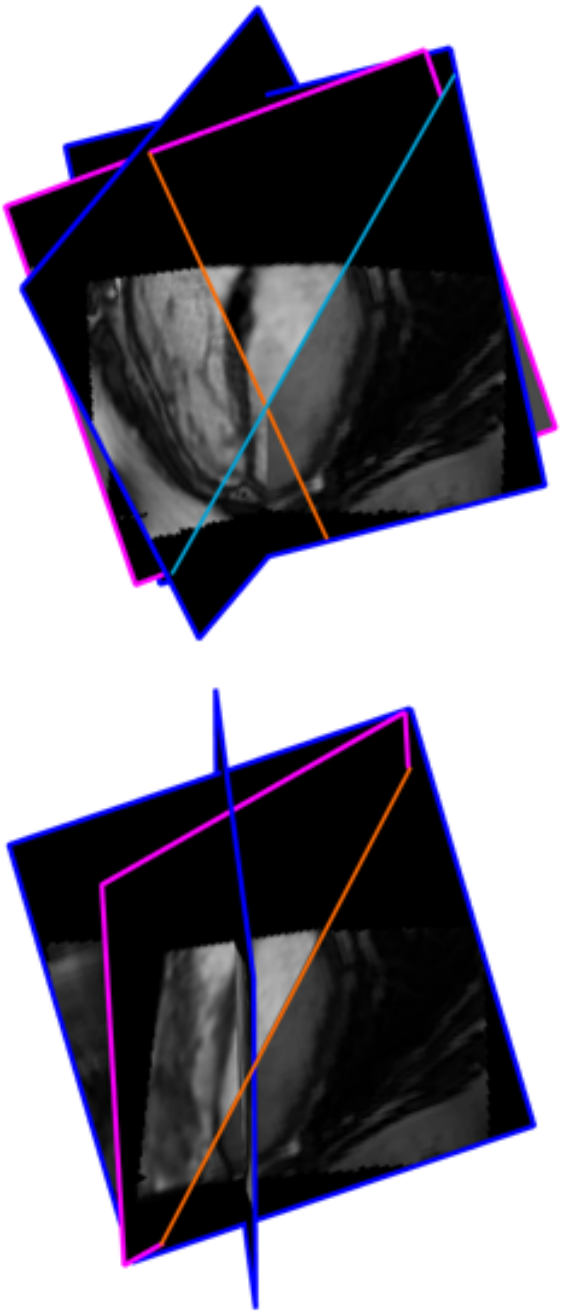


Figure 6.10: 3D views of the two original long axis slices (blue) and a 3rd additionally reconstructed slice (pink), and a different view (bottom) showing intersection of the myocardium of the left ventricle. The orange and blue lines indicate where the planes are intersecting, the profiles along those intersections are shown in Fig. 6.11.

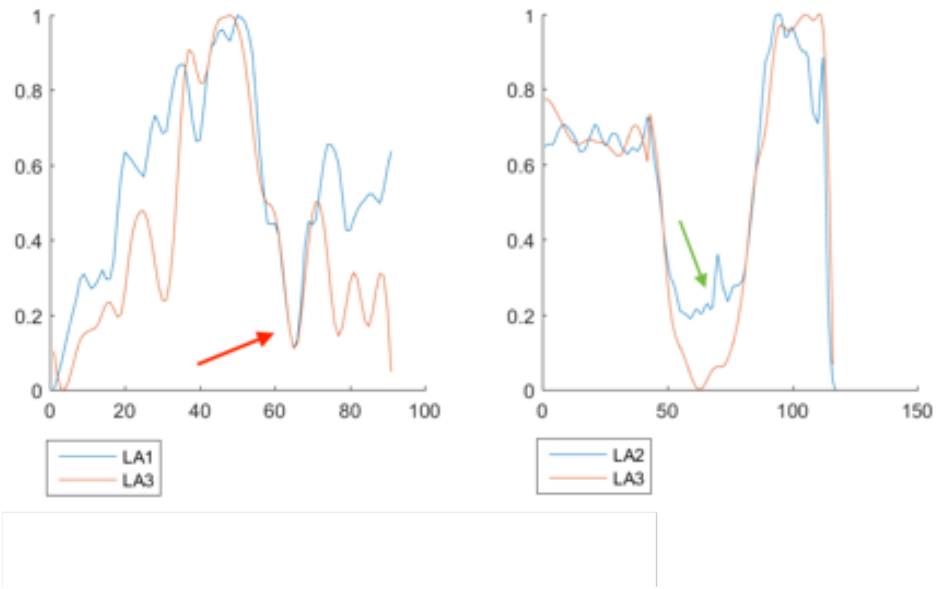


Figure 6.11: Profiles of the intersections of the true long axis slices with the additional reconstructed long axis slice shown in Fig. 6.10. The orange plot represents the profile along the 3rd long axis slice at the location where it overlaps with the true slices, whereas the blue shows the profile along the ground truth. An arrow is pointing to a dip in both plots, which corresponds to the myocardium.

networks will have different abilities to represent the data.

For our first attempt of implementing the original SRCNN architecture, we used the interpolated short axis data. The results were not promising: the output images were overly smooth images preserving approximate shapes of the ventricles, but not beyond that. An example is shown in Fig. 6.12. This comes as no surprise, as the MSE promotes smoothness as the best prediction for cases where the input and the label do not match spatially, nor in intensity values. This prompted the use of synthetic data, as it has been done in the relevant literature (e.g. [84]), for the first steps. It quickly made us consider incorporating conditional GANs, as we were aware they do not rely solely on pixel-to-pixel losses.

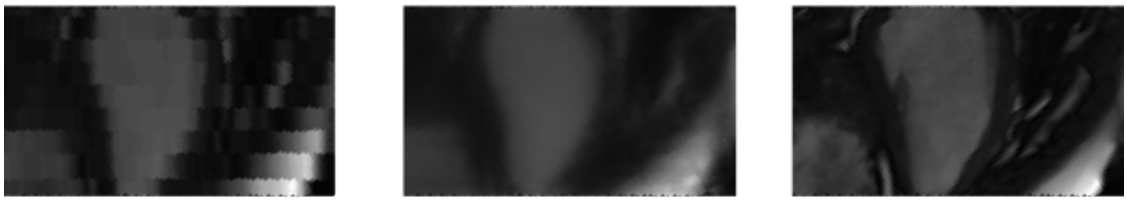


Figure 6.12: Result of the original SRCNN architecture trained with combined short and long axis data.

The results on the SRCNN architecture are not necessarily a fair comparison because of the large difference in scale between our aim and SRCNN [116], which aims to isotropically up-scale images by a scale of 3 or 4 whereas we work with highly anisotropic data. We have also started looking into residual learning, as several such applications showed good results [84, 120]. When applied to the SRCNN based architecture, improvements are noticeable when using exactly the same setup. We show an example of results obtained using residual learning in Fig. 6.13, where improvements are visible comparing the residual approach to the initial approach. It is possible that, instead of applying a larger kernel size in SRCNN as we have done, it could make more sense to use two smaller kernels before the activation function. Using two layers with small kernels results in a larger receptive field, but that would change the reasoning behind SRCNN and its simple 3-layer architecture.

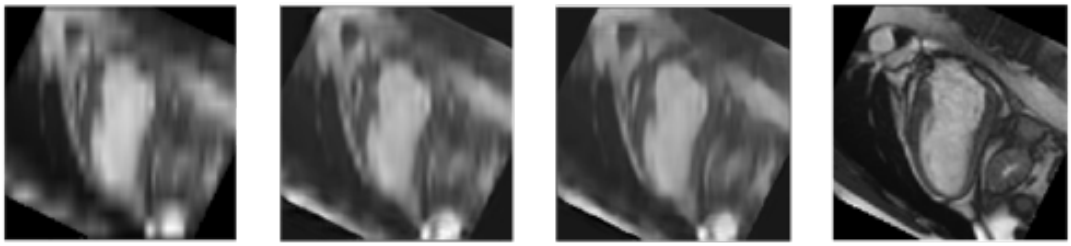


Figure 6.13: Interpolation, result of $SRCNN_{LA}$, $SRCNN_{LA}$ with residual learning, ground truth.

We show quantitative results for PSNR and SSIM for synthetic data in Table 6.2. We see the best performance is obtained by $U - net_{LA3}$, closely followed by $U - net_{LA4}$ and LAGAN, and outperforming $SRCNN_{LA}$ and the interpolations by several dB in terms of the PSNR. The SSIM is reportedly the highest by a small margin for the results from the LAGAN model. These numbers correlate with the results shown in both Fig. 6.5, showing results on synthetic data for all learned models and interpolation for four different hearts, as well as in Fig. 6.6, which shows a comparison of interpolation, $SRCNN_{LA}$, $U - net_{LA3}$ and the LAGAN models for a single long axis example. There are some limitations of using PSNR for quantifying the quality of the results. We know that PSNR can favour smooth images versus those with more texture, even if the latter is more similar visually to the ground truth. This is what happens in this case, as one could appreciate in Fig. 6.6 where the most visually accurate image is the result of the LAGAN model in the bottom right. From a qualitative point of view, it is hard to differentiate between both U-net models, they do show considerable improvements over interpolation and the $SRCNN_{LA}$ model seen applied to four different long axis views in Fig. 6.5. A noticeable difference between image appearance and the texture can be seen in Fig. 6.6, where the LAGAN model produces a result which appears almost in-differentiable to a real acquisition unlike the overly smoothed results provided by the U-nets. An interesting observation is that the vessels in the top of the image are not well recovered, whereas in Bhatia et al. [78] the synthetic results show that the vessel appears to be perfectly recovered, from the information contained in the equivalent of 2 short axis images a perfect circle is recovered. This could be because the training data contained the particular example shown in the results. Fig. 6.7 shows results for a 2-chamber view of a reconstruction using the short axis data. The misalignment is noticeable, especially when looking at the nearest neighbour interpolation in the top left. While the sharpness of the image is somewhat recovered by the U-net model, we can see that it directly translated inconsistencies in the input, such as the misalignment of the slices giving the myocardium an irregular shape. Using the U-net architecture gives much superior results than interpolation

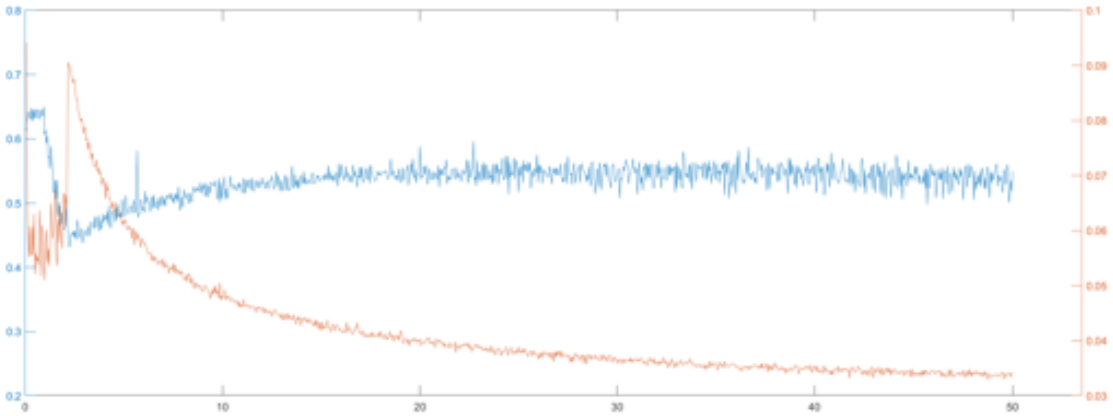


Figure 6.14: Discriminator (blue) and L1 (orange) losses for the SALAGAN, trained on short axis data.

and SRCNN, both quantitatively and in appearance. The resulting images, as shown in Fig. 6.5 are not blurry and the networks manage to recover missing features in structures as well as enhancing boundaries between blood pool and myocardium, which can aid further image analysis. When it comes to the different kernel sizes in the U-net, the smaller one produces marginally better quantitative results, as shown in Table 6.2. While qualitatively, there does not seem to be a difference between the two and the results look basically equal.

Moving to the results on short axis data, we show the quantitative results for the models trained on interpolated short axis stacks as input data in Table 6.3. Here we see the best performance for both the PSNR and the SSIM given by the SALAGAN model. For the results using conditional GANs, we show the discriminator loss and the L1 loss for training the SALAGAN model in Fig. 6.14. We note that the training times for this particular network exceeded 72 hours, due to the large parameter number and the amount of training data. We can see that the discriminator loss in blue starts at high confidence (65%) in its ability to differentiate between true and generated images. However after the 15^{th} epoch, the discriminator output starts to solidify just above the 50% line, indicating it is unable to differentiate the network

output from the ground truth labels anymore. On the other hand, we can see that the L1 loss in orange is decreasing, indicating the network output is not deviating too much from the input in a L1 sense. We want the low frequency information from the input to be preserved in the output. If the discriminator were to converge towards 0 or 1, it would indicate that the generator is not able to produce realistic enough output, or that the discriminator does not work.

In Tables 6.2 and 6.3, we can see that the quantitative analysis of our results is similar to the literature when it comes to the performance of GANs on the synthetic data. When it comes to the interpolated short axis data, the SALAGAN model outperforms the other methods both quantitatively and qualitatively. The average values of SSIM are higher for the GAN when compared to interpolation or the other networks trained on synthetic data. In the PSNR evaluation, the $U - net_{LA3}$ model outperforms the others, it may be noted that the error values associated with the GAN are much larger than for those of the other results, this may be due to the fact that the non-adversarial networks are trained to maximise the PSNR. This also correlates with some of the previous findings in the literature related to GANs used for super resolution. No improvements on PSNR have been observed in a number of studies [119, 122, 135], while other studies do show some improvements in that aspect [139, 141]. It is worth noting that latter group of studies mentioned makes use of some condition on the output, whereas the former group does not. This supports the idea behind conditional GANs, which put a constraint on the output having to match the input using a L1 correlation of the reconstructed output with the incomplete input. For the models trained on interpolated short axis input data, we observe that the SALAGAN outperforms all of the other models in terms of SSIM, as well as by several dBs in PSNR.

Interestingly, the Table 6.3 shows inferior results when it comes to PSNR for the $U - net_{SA3}$ and $U - net_{SA3}$ compared to all the other results. We can see in Fig. 6.7 that the intensities between interpolations and network results are different, especially when it comes to the blood pool. We can also see that that

the networks have amplified some structures such as the bright white stripe on the bottom left of the reconstructions, which are present in the interpolated short axis but not present in the ground truth image that is taken as a reference for PSNR evaluation. We can explain the inferior PSNR results as this metric works on a pixel-to-pixel level, and intensity and structural mismatch are penalised heavily. PSNR fails to describe the increase of sharpness of the myocardium. It is worth noting that this experiment was using misaligned data and a pixel-to-pixel loss for training, and these result are what prompted us to explore GAN models. This result however underlines one of the challenges that comes with the use of deep learning for image reconstruction and which will have to be addressed by future research [157].

Fig. 6.10 shows two rotated views of two acquired long axis slices and a third long axis slice that we reconstructed with the T-GAN model, that was pre-trained on the entire training set and re-trained on that particular heart. The normalised profiles along intersecting planes are shown in Fig. 6.11. The orange plots represent the profile along the reconstructed image, whereas the blue plots represent the profile of the ground truth data. It can be seen that the orange profiles are more smooth, and on the right hand side the dip does not feature the peaks which represent the papillary muscles, shown by the green arrow. A red arrow is pointing to a dip in the profile, indicating the myocardium. We can see that there is overlap at the major structures of interest, showing adequate representations of the left ventricle. However, objects such as the papillary muscles, which are well defined in the acquired long axis images, may not be as well represented in the short axis acquisitions which act as basis for the reconstruction of additional long axis slices.

One of the drawbacks of our presented methods of long axis slice reconstruction, is that the models will not be able to reconstruct any arbitrarily oriented slice through the stack. They will only be able to reconstruct orthogonal ones, since the training data assumes orthogonality between short and long axis slices. A drawback from the non-GAN models trained on interpolated short axis input, are the intensity

differences as well as the misalignment within the short axis stack, and with long axis slices are carried over in the results in the SRCNN and U-net models. While the results look visually improved, they are too similar to the corrupt input images and carry over any inconsistencies such as intensity mismatch or misalignment.

We have presented work on 2D multi-slice cardiac MRI images using static frames and shown possible applications using several different network architectures. Implications of the results we present here are two-fold. First, clinicians could forgo the long axis acquisition and therefore decrease total scan time. Alternatively, additional non standard long axis slices could be reconstructed. Combined with the possible improvements presented in chapter 4 such as the effectiveness of different slice protocols, this method could pave the way for further development and could radically change the way cardiac MRI is acquired today. We mentioned that the networks presented in this chapter had large parameter numbers, which resulted in long training times and the necessity of much training data. This was due to the doubling of the filter numbers at every level down the U-net architecture. In chapter 7, we will also reduce the number of filters. One of the aspects that were overlooked in the work presented in this chapter, is that we ignore a lot of the data that we could be using. Indeed, we only incorporate fractions of the pixels present in each short axis acquisition, the equivalent coverage of the long axis slices intersection. In addition to only making use of the data corresponding to locations of intersection between the short axis slices and the long axis acquisitions, we have also dissociate adjacent frames from each other and are losing another potential valuable source of information by treating frames individually. This is why, for the work presented in chapter 7, we focused on reconciling some of the issues that were just mentioned and developed a method allowing us to use sequences instead of individual frames as input to the network.

7

Reconstruction of Long Axis Slice Sequences from Short Axis Stack Sequences Using Convolutional Neural Networks

In the previous chapter, established deep learning architectures and approaches to super resolution of static images were applied to cardiac MRI short axis data and concluded on generative training techniques to bypass issues related to clinical data. The following chapter presents a deep learning approach where the dynamic aspect of frames in cine MRI sequences are used together via the introduction of long short-term memory layers into a well established network architecture. The aim is to introduce context sensitivity while also investigating into networks and smaller numbers of parameter with shorter training times. This work has been presented as an oral presentation at the RAMBO workshop at MICCAI 2018 in Granada, Spain [10].

7.1 Introduction

As we have outlined in chapter 5, machine learning methods have dominated the research in the biomedical image analysis field in recent years. With the increasing availability of computing power, large labelled datasets and open source libraries, deep learning has quickly become the benchmark for many tasks such as image classification and segmentation. SRCNN was the first application of deep learning to image super resolution and consisted of a simple network with three layers, inspired by the idea behind dictionary learning applications to super resolution. The first layer has a small filter size similar to a low resolution dictionary extracting a small low resolution image patch, the third layer a larger filter size similar to a high resolution dictionary upsampling to a bigger higher resolution patch, and the middle layer introduces a non linear mapping between the two [110]. A small number of training images underlines the simplicity of the approach, which shows pleasing results on natural images, and has become a benchmark in deep learning super resolution.

The new state of the art in image segmentation has been pushed by the U-net architecture, introduced by Ronneberger *et al.* [125]. The U-net is a convolutional neural network similar to an autoencoder including skip connections between input and output layers. The skip connections allow high frequency as well as low frequency information to be processed and make it suitable for super resolution, for which it has been applied to 3D microscopy in recent studies [126, 127]. Deep learning has also been applied to super resolution of cardiac MRI in [84], where a single image and a multi-image network are trained to predict residuals which are added to the low resolution image and give it high frequency information. The data used in that study comes from synthetically down-sampled 3D b-SFFP acquisitions that do not require realignment to account for breathing between acquisitions or patient motion. The same group recently extended the network to an anatomically constrained neural network that resembles a U-net and is able to do super resolution and segmentation aided by the addition of shape priors [88]. In contrast, our work

aims to improve standard dynamic 2D data acquired in clinical practice, using the dynamic information in the time sequence to improve the reconstruction. We present a network that learns a one-to-one mapping between low resolution and high resolution 2D image sequences to generate additional high resolution long axis view sequences from a dynamic short axis stack.

RNNs and especially LSTM networks are beginning to see applications in medical image analysis. Recurrence can be applied in a spatial sense, by considering adjacent slices in a 3D image. In a study on prostate MRI for cancer segmentation, adjacent 2D slices were fed into a U-net fitted with recurrent layers at every convolution [147]. RNNs have been applied to cardiac cine MRI first by Poudel *et al.* [148], at the lowest resolution level of a U-net to take advantage of low frequency features in consecutive frames of the cardiac cycle. LSTMs have been applied to enhance performance of myocardium segmentations in cardiac cine MRI sequences [150]. In that study, similar to [148], recurrent layers are present in the lower resolution levels of the network architecture. Up to our knowledge, convolutional LSTM networks have not yet been used for cardiac cine MRI sequence super resolution reconstructions, some work has been done for a compressed sensing approach using RNNs as described in [149]. In this chapter we propose a method that exploits the temporal aspect of cardiac cine MRI to recover high resolution long axis slice sequences from low resolution short axis stack sequences.

7.2 Materials and Methods

The method we propose uses an architecture inspired by the U-net with added recurrent layers, sharing some characteristics with those used in [147], [148] and [150] for segmentation. While [147] used recurrence on all levels of a U-net, [150] only on the two lower resolution levels, and [148] only on the lowest resolution level, we included recurrence layers on the first two layers, corresponding to those where the high frequency features are present in the images. Fig. 7.1 shows four subsequent isotropic decimations of a 128×128 image, representative of the input sizes at

different levels of the U-net we are using. We limited the number of levels to the first two for two reasons: to save memory and because unlike with segmentation work where the lower frequency features are more important, we are particularly interested in the high frequency information which is needed to convert low resolution into high resolution images.

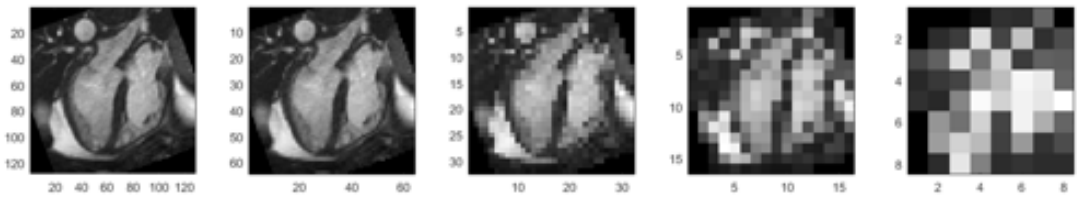


Figure 7.1: Representations of image size scale among the different levels of a U-net. A 128×128 image and its four subsequent down-sampled versions. Each following image has an isotropic decimation factor of 2, resulting in sizes 64×64, 32×32, 16×16, and 8×8.

Fig. 7.2 shows the network architecture. The network we propose is inspired by the U-net with a contractive part and skip connections sensitive to low and high frequency details, respectively. At the first and second levels, we introduced LSTM convolution layers. There are a total of five levels in the network each initiated by a 2×2 maximum pooling layer. The input data has a size of $128 \times 128 \times 10$, consisting of sequences of 10 frames each of size 128×128. The lowest level of the U-net therefore operates on samples of size $8 \times 8 \times 10$ where only the very low frequency features are present. This is illustrated in image 7.1, where no features recognisable to the human eye are left in the smallest sized image.

We chose to put the recurrent layers on the top levels since we want to enhance the high frequency features of the images and they are mostly present in the first and second second levels. Going down to the third level, the data is now $32 \times 32 \times 10$

which we have deemed to be less relevant to the high frequency content. We have also trained a U-net with the same architecture where all the convolutional layers are not recurrent, with filters of the same size 3×3, to compare performance between a context sensitive and a static network. Training was performed using an Adam optimiser to minimise mean squared error over 90 epochs, with a learning rate of 4.5×10^{-5}. The network was written in Python using the Keras library [153] running on Tensorflow backend, and training was performed on a Nvidia GeForce GTX 1080 Ti 256 RAM GPU.

In contrast to the U-nets in chapter 6, we have drastically decreased the number of filters. Instead of doubling the numbers when going down the layers of the U-net, we decided to decrease the filter numbers. With the aim of a lightweight network, we chose to cut down the filter numbers in the lower levels of the U-net, for the same reason as above since we focus on the high resolution features. The final architecture we used was as follows: $CL_{128} - M - CL_{64} - M - C_{32} - M - C_{32} - M - C_{32} - U - C_{32} - U - C_{32} - U - CL_{64} - U - CL_{128}$ where CL corresponds to a convolutional LSTM layer followed by a convolution, followed by a ReLU activation, C two convolutions followed by a ReLU, M a 2×2 maximum pooling operation and U a 2×2 up-sampling layer. The kernel size we used throughout this architecture was 3×3, as it results in lower parameter numbers and the difference between the two U-nets in the previous chapter was not significant. Compared to the parameter heavy U-nets and GANs from the previous chapter, the total number of parameters went down from $\mathcal{O}(10^7)$ to $\mathcal{O}(10^5)$ total parameters. The training times drastically went down as well, the training took about 5h compared to more than 3-4 days for the previous models such as SALAGAN.

7.2.1 Data

Long axis views from the Kaggle Data Science Bowl Cardiac Challenge Data [152] were used in training. The dataset consists of cine MRI sequences of over five hundred patients. Every data set comes as a cine movie of 30 frames, however the

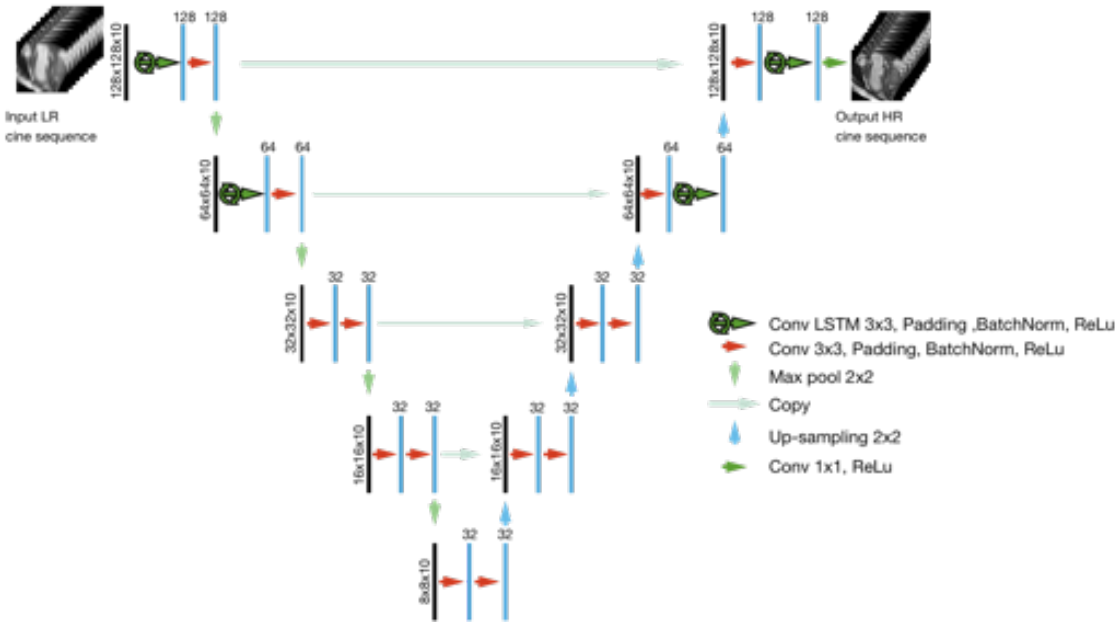


Figure 7.2: Network architecture. A network inspired by the U-net with four convolutional LSTM layers at the top of the network aiming to enhance high frequency features by taking into account the dynamic aspect of the cine sequence on the original image size and after the first max pooling operation which decreases the image size by a factor of two. Filter numbers are indicated on the diagram on top of the blue bars after every convolution. Arrows represent the different operations. Height, width, and time sizes are shown for every level.

number of short axis and long axis slices differ, as a standardised imaging protocol was not in place. VLA and HLA acquisitions were present for most of the patients but a non negligible part of the data had only one or no long axis views. Data missing long axis views as well as other unusable data, such as those affected by strong artifacts, have been discarded by visual inspection. The patients have a large spread of age and size which is advantageous to preserve the generalisation properties of the method, just as in chapter 6. All remaining images were resampled to isotropic resolution of 1.4mm x 1.4mm, rotated to the same upright orientation

where the base is located towards the top of the image and the apex towards the bottom. We then down-sampled the images in the baso-apical direction, in order to match the slice thickness of standard short axis slices of 10mm. In this way, we generated images similar to those that would be represented by the short axis stack. Every sequence was also normalised such that all pixel intensities lie in the range between 0 and 1. We did not differentiate between HLA and VLA views, we have included them both together in the training, validation, and testing datasets. In this way, we aimed to demonstrate the ability of the method to recover images with different appearance (e.g. in terms of the number of chambers), with the eventual goal of using the network to produce reconstructions of slices in any arbitrary orthogonal orientation from short axis stacks.

After splitting the sequences of 30 frames into shorter sequences of 10 frames each (to reduce the time needed for training as well as increasing the amount of training data), 3342 low-high resolution sequence pairs of 10 frames per sequence were available. 3000 sequences were used for training, 171 set aside for validation, and 171 for testing, ensuring that none of the split sequences were spanning over the training and the validation or testing set. We have also ensured that no two sequences (e.g. VLA and HLA) from the same subject were dissociated between the training, testing, and validation sets. For the static network, all the frames in a sequence were used, which increased the training, validation, and testing data by a factor of 30, just as in chapter 6.

7.3 Results

	Interpolated	**static U-net**	**LSTM U-net**	$U-net_{LA3}$*	LAGAN*
PSNR	23.17dB	25.23dB	26.57 dB	26.10dB	25.38dB
SSIM	0.72	0.77	0.81	0.799	0.804

Table 7.1: Quantitative evaluation (PSNR and SSIM) of interpolated, single frame U-net, and the proposed network results on the whole testing data set which has not been seen by the networks in training. Models from chapter 6 are also included in this table and are denoted by a star, they carry a higher number of parameters.

Results on the first 5 frames of a HLA sequence are shown in Fig. 7.3. A representative result on 5 non-adjacent frames of a VLA sequence can be seen in Fig. 7.4, with the cardiac contraction more easily visible due to the frames spanning a longer time. Both figures display a sequence from the unseen testing dataset using cubic interpolation in the first row, the result of using static frames only in the second row, the result of the proposed network in the third row, and the ground truth on the bottom row. Each frame has been magnified around the apex, one of the features that is most prone to being missed by the short axis stack acquisition. The proposed network manages to recover the apex across the sequence, with much better definition than previously used standard U-Nets.

In addition to qualitative improvements, the PSNR and SSIM of the dynamic network output outperform the static network and interpolation. Quantitative results are shown in Table 7.1, which contains the average values for PSNR and SSIM of the whole testing data set and shows that the dynamic network output is superior to the static network as well as interpolation. The table also contains values from chapter 6, showing that the heavyweight static networks are performing better than the static lightweight network, and slightly worse than the dynamic lightweight network.

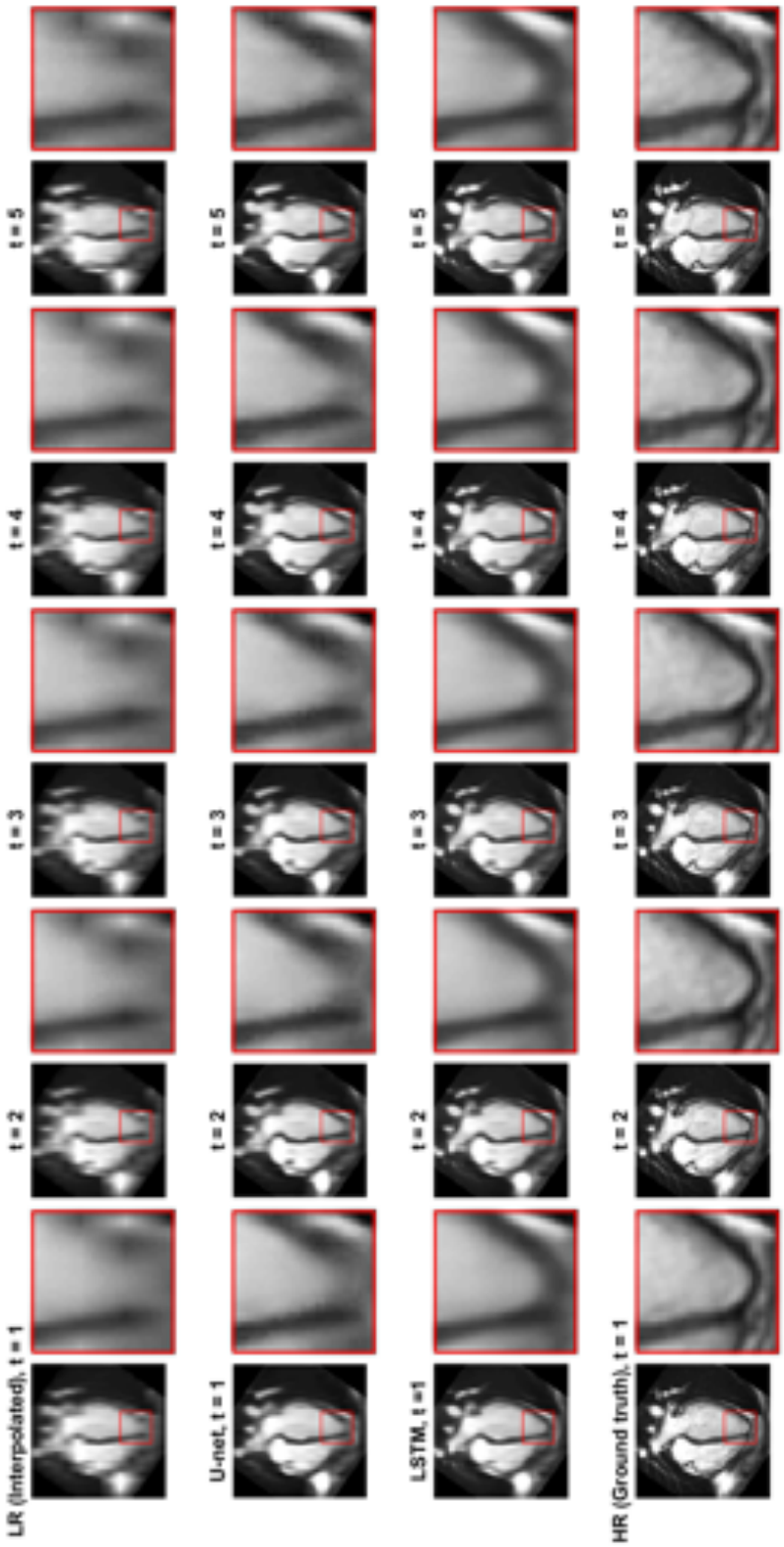


Figure 7.3: Result on the first 5 frames of a HLA view cine sequence. The top row shows the low resolution interpolated input, the second row shows the result given by the static U-net, the third the result given by network including LSTM layers, and the bottom row shows the high resolution ground truth. This proposed enhanced 4-chamber sequence has a PSNR of 25.28dB and a SSIM of 0.82 while the static U-net gives a PSNR of 23.83dB and a SSIM of 0.79 and the interpolated sequence a PSNR of 22.15dB and a SSIM of 0.73.

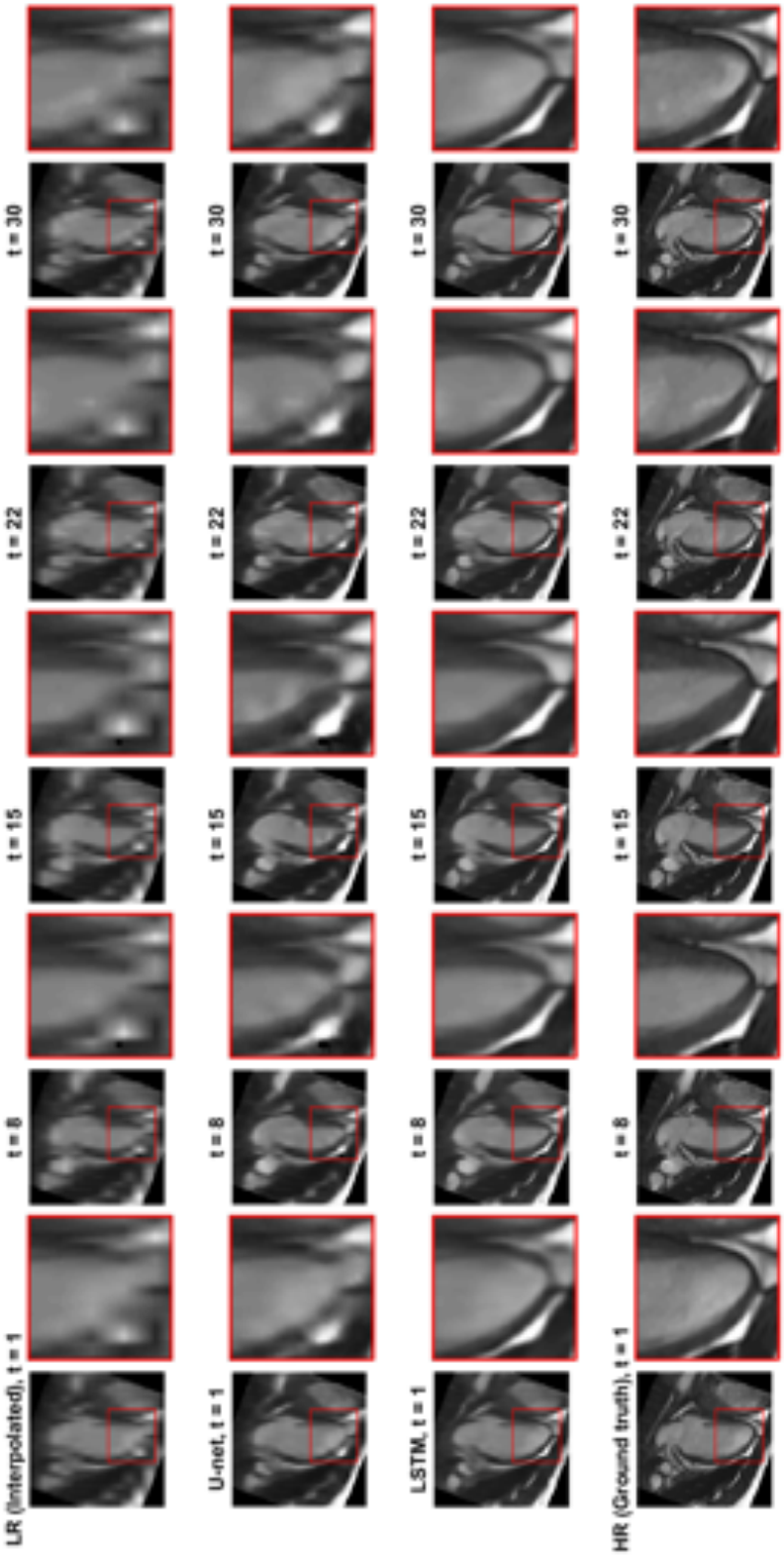


Figure 7.4: Result on the 1st, 8th, 15th, 22nd, and 30th frames of a VLA view cine sequence. The top row shows shows the low resolution interpolated input, the second row shows the result given by the static U-net, the third the result given by network including LSTM layers, and the bottom row shows the high resolution ground truth. This proposed enhanced 2-chamber sequence has a PSNR of 27.55dB and a SSIM of 0.83 while the static U-net gives a PSNR of 25.69dB and a SSIM of 0.78 and the interpolated sequence a PSNR of 22.23dB and a SSIM of 0.71.

7.4 Discussion and Conclusion

We have shown that there is an advantage to using the temporal context for super resolution of cardiac cine MRI sequences, in comparison to the more common approach of reconstructing individual frames. Our proposed architecture, which includes LSTM layers on the two upper layers of a U-net, gives qualitatively and quantitatively superior results to an equivalent U-net architecture with no recurrence.

In order to concentrate on the effects of recurrence on reconstructions, we chose experiments that avoid common acquisition artifacts such as misalignment between slices and intensity mismatches. Future work will look into the reconstruction from clinical short axis stacks suffering from these artifacts, as well as reconstruction of arbitrarily oriented slices, eventually aiming to reconstruct complete 3D datasets.

This also means that in the data we have been using, full coverage of the long axis is achieved with the synthetic short axis equivalent. Features such as the base and the aorta on top of the image would not normally appear in short axis images because the stack coverage usually will not go this high. During the paper discussion at the RAMBO workshop, a member of the audience expressed his appreciation that the atrioventricular displacement might be reduced by the inclusion of the recurrent network in our method. This particular displacement has been shown to be one of the main driving forces of stroke volume generation [158]. Recurrent layers can have the effect of smoothing in the time dimension, which could be a reason for this remark. We subsequently decided to review this aspect in a more accurate way. In Fig. 7.5 we show the end diastolic and the end systolic frames for our reconstruction as well as the ground truth equivalents, superimposed with guiding lines showing the heights of the atrioventricular displacements. The lines were drawn on the ground truth images and transferred to the images displaying our result. It can be indeed argued that one may notice a slight shortening in the bottom left frame, on the basis of an apparent fading out of the intensity and therefore lack of detail of the myocardium. This is especially noticeable in the

zoomed region of the end diastolic frames of Fig. 7.5, it is not as noticeable in the end systolic frames. The extent of the loss of detail is also due to the L1 loss function which favours smooth appearance. While we believe that the small size of this effect is unlikely to produce significant changes in diagnosis, this might require further research in specific applications.

A video can be viewed under this hyperlink `https://imgur.com/a/04w2px8`. In future work, the size of the kernels in convolutional LSTM layers can be explored, since different sizes can capture different motion speeds. We have selected 3×3 on the basis that the previous U-net we have trained also used the same sizes, and promising results were obtained with higher filter numbers. In this work, we have followed the so-called many-to-many approach. This means that the number of input and output samples in the sequence are the same. Other variants are possible, including many-to-one where the output to an input sequence is a single datapoint, or frame. Another approach could be to make a network that takes the entire sequence as its input, and outputs the end diastolic and end systolic frames, take advantage of the contextual information from neighbouring frames and to perform analysis on those two frames, which may be used for ejection fraction calculation. Future work could start with a combination the work of chapters 6 and 7, to investigate recurrent GAN models in order to move ahead to make use of acquired short axis data and its temporal context. We show results in 5 frames of a sequence of interpolated short axis data in Fig. 7.6, where some limitations we tried to overcome with the synthetic data become apparent. The top row shows the interpolated short axis data, the second the result using the static U-net, the third row the LSTM model, and the ground truth at the bottom. The results from the LSTM model achieve some improvement in the image appearance compared the the interpolation and the static U-net. We can also see that the intensity inconsistencies are carried over in both results obtained from the networks, and the incomplete apex does not get fully recovered.

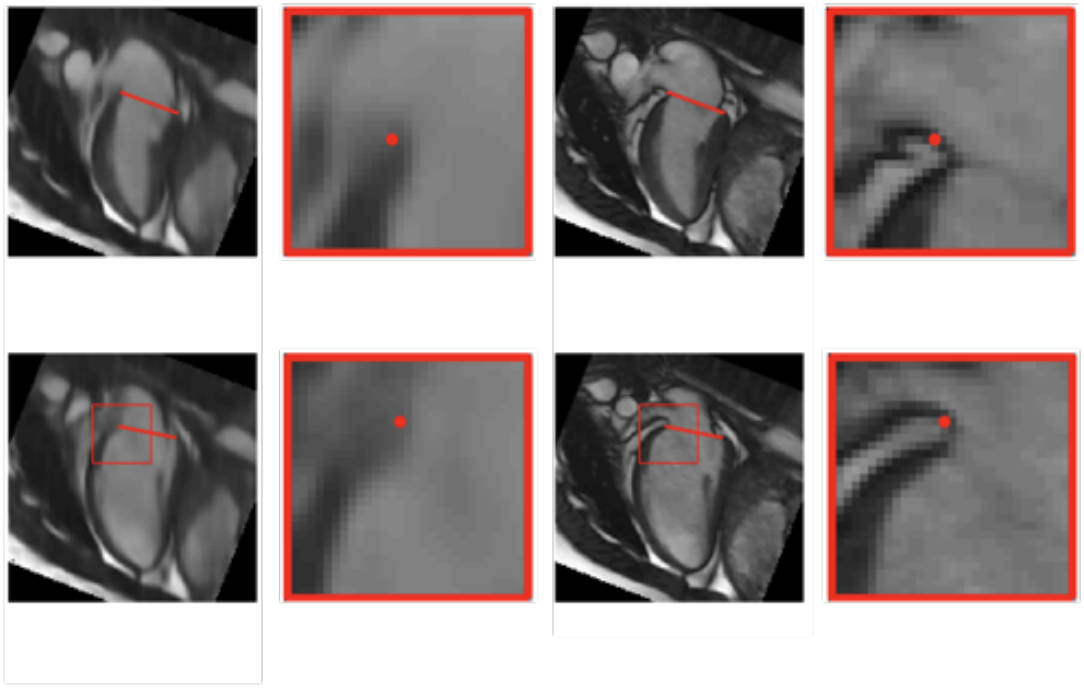


Figure 7.5: Top: End systole, when the ventricles are at the end of contraction and display the lowest blood volume of the sequence, bottom: end diastole, when the ventricles are at their maximum volumes and about to contract. Red boxes represent zoomed regions. Red dots are present at the same locations as reference points at the end of the myocardium to display the shortening.

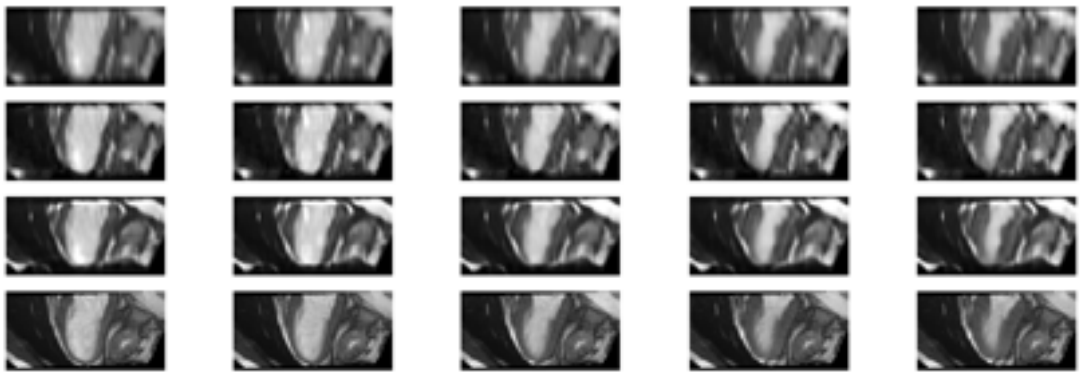


Figure 7.6: Top: interpolated short axis, 2nd row: static U-net result, result on LSTM model, ground truth.

8

Conclusions and further directions

The previous chapter concluded the last analysis contribution of the thesis and explained work on super resolution of dynamic cine MRI sequences using convolutional LSTM networks. This final chapter is used to summarise the thesis and point to future research directions.

Discussion and Conclusions

Multi-slice cardiac cine MRI acquisition protocols were born out of a compromise aiming to get the best coverage of the heart in a time efficient and reproducible manner. Literature on post processing of cardiac MRI highlights the differences between theory and practice, where most of the literature on cardiac MRI super resolution reconstruction treats either slices maximising the coverage of the heart and overlap, or ideal synthetic data without regards to the real problem that has to be overcome with commonly available data. This thesis, too, was a result of compromise. It started out as a project on dictionary learning, based on large databases. We wanted to combine cutting edge methodology and the data acquired in practice, since it was missing in the literature. It became apparent that this was for good reason: misalignment and inconsistencies between the slices made it impossible, and led us to drop dictionary learning and investigate different slice protocols, with the difference to the majority of the literature that we did not require additional scanner time by increasing the number of acquisitions per patient. When moving to the next part of the project, a waiting period for UK Biobank data (we eventually got access at the beginning of the 2018 academic year) prompted us to resort to the publicly available Kaggle data.

In chapters 2 and 3, we have presented background on MRI, and particularly cardiac MRI to the reader as well as background and a review of the literature on super resolution with a focus on traditional techniques applied to MRI and cardiac MRI. We have seen, that because of some of the drawbacks related to MRI acquisitions such as the long acquisition times or motion, something which is amplified in cardiac MRI, special acquisition protocols have been designed to cover the heart by multi-slice imaging when isotropic high resolution and 3D volumes with high signal-to-noise ratios are difficult to achieve. In addition to that, efforts to reduce imaging times via compressed sensing for smart sampling of the k-space, and at post-processing have been sought after and have generated a substantial amount of literature which is still growing to this day, showing it is still a very active area

of research right now. High resolution coverage of the heart resulting in the ability to generate 3D meshes to diagnose and monitor cardiac anatomy and function is of high importance. When performing repeated acquisitions of the heart of a same patient using multi-slice protocols, it is unlikely that the same exact planes are going to be imaged. Therefore, for reproducibility, it is important to be able to capture the heart in 3D more efficiently, and is an additional reason why research in this direction matters. We have seen that a majority of the research in cardiac MRI super resolution [66, 67, 69, 71, 84] deals with data which is not necessarily the kind of data that would be acquired in clinical practice. While results are impressive, the data used is not representative of data acquired in clinical practice. Protocols with additional slices would require more scanner time, when the aim of such methods is to avoid just that.

In chapter 4, we have presented a method which combines short and long axis slices in a contrast independent manner to perform 3D reconstructions. The method includes a new application of 3D directional total variation. In addition to that, we propose a slice protocol with a balanced number of short and long axis slices, resulting in the same total number of acquired slices and that therefore would not drastically change the scanner time required. We show that balancing the number of short and long axis slices results in improved reconstructions. This is in agreement with the literature on super resolution in MRI where overlap between the slices, which is minimal in clinical cardiac multi-slice MRI acquisitions, is favoured. We have performed validation on pre-clinical data of excised rat hearts and shown it also works on human in vivo data where we generated short and long axis images with differences in contrast, from 3D volumes. Because of the way clinical multi-slice cardiac MRI is laid out, our method is not able to treat that kind of data that does not seek a maximisation of overlap between slices, in addition to the fact that misalignment between slices makes it difficult.

Because of this drawback preventing us from dealing with data following standardised protocols such as the UK Biobank [35], as well as the recent surge in methods utilising deep learning, we chose to change the focus of our research to those methods. In chapter 5, we present the reader with a background on deep learning methods, especially CNNs, and their applications to super resolution and medical image analysis. Since the rediscovery of deep learning techniques, driven by the availability of large labeled datasets, higher computing power, improvements in methodology, and user friendly libraries, an ever growing amount of research contributions and mainstream interest has surged. It has not taken long for deep learning, and more specifically CNNs, to be applied to super resolution [110]. Another milestone was the U-net [125], which has become state of the art for segmentation and a main component of many subsequent image analysis efforts [88, 126, 127], noticeably in conditional GANs such as one proposed for image to image translation [134].

In chapter 6, we present work using static images which had the aim of generating long axis slices from short axis stacks. We first dealt with synthetic data generated by down-sampling long axis slices only. We compared results obtained from interpolations, adaptations of SRCNN and U-nets, as well as conditional GANs, and fine tuned a conditional GAN using transfer learning. Qualitative and quantitative results showed improvements of: SRCNN over interpolations, U-nets over SRCNN, and of conditional GANs over U-nets. We then moved to using real short axis data, affected by slice misalignment and intensity inconsistencies. Results from SRCNN and U-nets showed no real improvements in analysability of the images, whereas conditional GANs gave promising results, and showed the ability to recover incomplete features that may not be fully captures by the short axis stack. One of the main holdbacks from this kind of 2D slice generation from short axis stacks, is that we are leaving out all of the out-of-plane information of short axis slices, as well as the temporal information from the cine MRI sequences. Therefore, we aimed to address the issue in the next chapter.

In chapter 7, we have presented our work on super resolution of dynamic cine sequences using a combination of a U-net architecture with added convolutional LSTM layers, which add a recurrence element allowing to exploit the temporal aspect of cine MRI data. We show that using a lightweight architecture, with fewer parameters than the U-nets used in chapter 6, the incorporation of the dynamic aspect of the data allows for enhanced reconstructions when compared to the same lightweight architecture lacking the recurrence element. As opposed to approaches to segmentation in the literature using recurrence in CNNs at the lowest layers of a U-net [148,150] or at every level [147], we chose to incorporate the LSTM layers only at the upper levels of a U-net, where high frequency information is mostly present. Segmentation approaches, especially ones for left ventricle in cardiac MRI, mainly draw information from the low frequency information, which contains most of the global image appearance which is necessary for adequate results. For super resolution reconstruction, we aim to recover the high frequency information so it made sense to apply the recurrence to the top layers. Our results on synthetically generated low resolution sequences show promising results in their ability to recover important features of the heart, qualitative as well as quantitative evaluation with PSNR and SSIM outperforms interpolations and static U-nets. Over-smoothing of some features in the features at the base is present when looking at the video sequence and worth looking into for further research, even though these features will not normally be covered by the short axis stack.

Traditional methods based on MAP reconstructions such as the one we presented in chapter 4 are interesting because they do not rely on a big database or any kind of learning. Those approaches are especially important because large databases are more difficult to come by when it comes to medical data, and even more so for unhealthy populations. However, such traditional MAP methods require specific imaging protocols that are designed to maximise overlap between intersecting planes, which is not the case for the cardiac imaging protocols followed in clinical practice. Deep learning based methods have been outperforming standard methodologies in

most computer vision related tasks and they are set to stay in the spotlight, given the current trend of research output. Our contributions from the regularised least squares and the deep learning chapters could be combined via a novel directional total variation loss function for deep learning. Conventional total variation has been used as a loss function in deep learning applications before in [159], however the directional total variation would allow an additional constraint on the shape or structure of the label and the input and not just a loss promoting smoothness. We have seen that shape constraints for deep learning approaches have yielded superior results for super resolution and segmentation of cardiac MRI, in the previously mentioned [88].

In order to achieve a better sense of validation for the work in chapter 4, we would require acquisition of multi-slice clinical data with the proposed protocol of balanced short axis and long axis slices in the future. Combined with results from chapter 4, where we show the validity of alternative slice protocols, our deep learning work and the current research climate in general could radically change the way cardiac MRI is acquired today.

Future directions

The research we have presented in this thesis and the growing amount of literature in the field are pointing at the fact that we have arrived at a stage in the research, where good quality results are achieved by deep learning when it comes to reconstructions using multi-slice 2D cardiac MRI as currently acquired in clinical practice. We have shown it is feasible to reconstruct long axis slices from short axis stacks, and other studies have shown promise with up-sampling short axis stacks to 3D volumes [84, 88]. The quality of the results so far is only assessed and validated by image reconstruction metrics, or through image analysis such as the error of segmentations. It is in the clinical aspect where the difference is yet to be made so

now the next steps will be to move from satisfying engineering metrics to impactful changes in clinical practice. Real impact will be achieved when imaging times are reduced and the imaging and analysis reproducibility are facilitated by implementing updated image acquisition protocols that match the data used in the literature. We have shown in Chapter 4, that a balanced number of long and short axis slices allow for good reconstruction based on regularised least squares and may be worth further research, and seeing whether it is realistically implementable in practice. In addition to that, the recent study by [75] also supports this concept of balancing short and long axis slices, the authors show that segmenting balanced sparse slices as opposed to the usual short axis stack is advantageous for segmentation and subsequent mesh generations. We believe that deep learning is going to drastically change the way routine clinical cardiac MRI exams are done over the next decade, based on the work done on both ends of the imaging pipeline with compressed sensing and super resolution reconstructions. Other aspects of the imaging process have also been investigated using deep learning, such as in a recent study exploring automatic evaluation of the short axis data coverage [160]. With approaches like this, automation and optimisation of cardiac MRI acquisition protocols could take impactful time-saving strides in the coming years. Before being able to start implementing deep learning based techniques into clinical practice, there are a number of aspects to be addressed. Work will need to be validated with a larger population, including a number of abnormalities, to ensure that the networks do not have a bias to the representation of healthy anatomy. We now have access to the more homogeneous UK biobank data, so it will be worth looking into repeating our experiment from Chapter 6 with this more coherent data and seeing wether fine tuning the network to the Kaggle data we have used produces worthwhile results. As shown by Bai *et al.* [156], fine tuning a network trained on homogeneous data to extend the network to match slightly different data is a viable option. As the authors mentioned in their paper, fine tuning will always require labeled data, which may not always be easy to come by. It also has to be established, which imaging protocol will be implemented for deep learning based reconstructions in the future.

Deep learning based on data comprising of very few examples is likely to be sought after, because of problems such as the scarcity of labeled medical data, as well as the difficulty of generalising patient data when it comes to disease. One major improvement to our work would be to use the full volume covered by the short axis slices. By dissociating volumetric or temporal data, valuable contextual information gets lost. 3D U-nets have been applied to the previously mentioned microscopy super resolution contributions [126,127]. A possible direct extension of our work could be to use the full interpolated short axis volume, coupled with a volume containing a mask, in order to extract slices. To see whether it would improve performance, it would also be interesting to implement residual learning to our models, as shown in Fig. 6.13. Residual learning has shown clear improvements on SRCNN. As an alternative method to the conditional GANs, it would be interesting to implement a cycle GAN [161] with the same interpolated short axis data we have been using. Unlike the conditional GAN or most other networks, the training data for the cycle GAN is unpaired. To put it into our context, it would dissociate the interpolated short axis data from the long axis slices. Cycle GANs work by learning a mapping from the input data distribution to the output data distribution using adversarial loss, and constraining it to also work with the inverse mapping, which makes it cyclical, without pairing the individual input and output samples. During the training of cycle GAN models, no ground truth images are required as label. This could be a way of overcoming misalignment and intensity mismatch.

Bibliography

[1] Prachi Bhatnagar, Kremlin Wickramasinghe, Julianne Williams, Mike Rayner, and Nick Townsend. The epidemiology of cardiovascular disease in the UK 2014. *Heart (British Cardiac Society)*, 2015.

[2] Peter Scarborough and Peter Weissberg. *Trends in coronary heart disease, 1961-2011*. British Heart Foundation London, 2011.

[3] Albert de Roos and Charles B Higgins. Cardiac radiology: Centenary review. *Radiology*, 273(2S):S142–S159, 2014.

[4] Frank Grothues, Gillian C Smith, James CC Moon, Nicholas G Bellenger, Peter Collins, Helmut U Klein, and Dudley J Pennell. Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy. *The American journal of cardiology*, 90(1):29–34, 2002.

[5] Frank Grothues, James C Moon, Nicholas G Bellenger, Gillian S Smith, Helmut U Klein, and Dudley J Pennell. Interstudy reproducibility of right ventricular volumes, function, and mass with cardiovascular magnetic resonance. *American heart journal*, 147(2):218–223, 2004.

[6] J Paul Finn, Kambiz Nael, Vibhas Deshpande, Osman Ratib, and Gerhard Laub. Cardiac MR imaging: State of the technology 1. *Radiology*, 241(2):338–354, 2006.

[7] Wufeng Xue, Andrea Lum, Ashley Mercado, Mark Landis, James Warrington, and Shuo Li. Full quantification of left ventricle via deep multitask learning

network respecting intra-and inter-task relatedness. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 276–284. Springer, 2017.

[8] Wufeng Xue, Gary Brahm, Sachin Pandey, Stephanie Leung, and Shuo Li. Full left ventricle quantification via deep multitask relationships learning. *Medical image analysis*, 43:54–65, 2018.

[9] Nicolas Basty, Darryl McClymont, Irvin Teh, Jürgen E Schneider, and Vicente Grau. Reconstruction of 3D Cardiac MR Images from 2D Slices using Directional Total Variation. In *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*, pages 127–135. Springer, 2017.

[10] Nicolas Basty and Vicente Grau. Super resolution of cardiac cine MRI sequences using deep learning. In *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pages 23–31. Springer, 2018.

[11] Gillian Pocock, Christopher D Richards, and David A Richards. *Human physiology*. Oxford university press, 2013.

[12] *Cardiac anatomy*, (accessed June 6, 2016). http://www.texasheart.org/HIC/Anatomy/images/fig1_crosslg.jpg.

[13] Heart disease and stroke statistics – 2009 update. Technical report, American Heart Association, Dallas, Texas, 2009.

[14] Yeo, Michael Papadakis, Sanjay Sharma, Cox, M Sheppard, and Elijah Behr. Young sudden cardiac death in England and Wales: a decade of change. *European Heart Journal*, 38(suppl_1), 2017.

[15] Diagnosing a heart attack, magnetic resonance imaging. https://www.heart.org/HEARTORG/Conditions/HeartAttack/DiagnosingaHeartAttack/Magnetic-Resonance-Imaging-MRI_UCM_441632_Article.jsp. 2015.

[16] Paul Leeson, Daniel Augustine, Harald Becher, and Andrew RJ Mitchell. Oxford specialist handbooks in cardiology: Echocardiography, 2012.

[17] Lawrence Boxt and Suhny Abbara. *Cardiac Imaging: The Requisites E-Book*. Elsevier Health Sciences, 2015.

[18] Vasken Dilsizian and Gerald M Pohost. *Cardiac CT, PET and MR*. John Wiley & Sons, 2011.

[19] Patrick J Scanlon, David P Faxon, Anne-Marie Audet, Blase Carabello, Gregory J Dehmer, Kim A Eagle, Ronald D Legako, Donald F Leon, John A Murray, Steven E Nissen, et al. ACC/AHA guidelines for coronary angiography: executive summary and recommendations: a report of the american college of cardiology/american heart association task force on practice guidelines (committee on coronary angiography) developed in collaboration with the society for cardiac angiography and interventions. *Circulation*, 99(17):2345–2357, 1999.

[20] Donald W McRobbie, Elizabeth A Moore, and Martin J Graves. *MRI from Picture to Proton*. Cambridge university press, 2017.

[21] Andres Kroker Keelin Rivard Soumya Sudhindra David Adair, Matthew Walton and Shefali Pandey. Physics of MRI. http://199.116.233.101/index.php/Physics_of_MRI. [Online; accessed 19-July-2018].

[22] David D Stark and William G Bradley. *Magnetic Resonance Imaging*. Mosby, 3rd edition, 1999.

[23] John P Ridgway. Cardiovascular magnetic resonance physics for clinicians: part i. *Journal of cardiovascular magnetic resonance*, 12(1):1, 2010.

[24] Anil Bharath. *Introductory Medical Imaging*. Morgan & Claypool Publishers, 2009.

[25] Ernesto Castillo, João Lima, and David Bluemke. Regional myocardial function: advances in MR imaging and analysis. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 23 Spec No:S127–S140, 2003.

[26] Sandra Pujadas, Gautham P. Reddy, Oliver Weber, Jennifer J. Lee, and Charles B. Higgins. MR imaging assessment of cardiac function. *Journal of Magnetic Resonance Imaging*, 19(6):789–799, 2004.

[27] Donald Longmore, Stephen Underwood, Godfrey and Hounsfield, Phillip Poole-Wilson, Denison, Richard Klipstein, David Firmin, Mikio Watanabe, Kim Fox, et al. Dimensional accuracy of magnetic resonance in studies of the heart. *The Lancet*, 325(8442):1360–1362, 1985.

[28] Michael A Guttman, Elias A Zerhouni, and Elliot R McVeigh. Analysis of cardiac function from MR images. *Computer Graphics and Applications, IEEE*, 17(1):30–38, 1997.

[29] Peter MT Pattynama, Albert De Roos, Ernst E Van der Wall, and Ad E Van Voorthuisen. Evaluation of cardiac function with magnetic resonance imaging. *American heart journal*, 128(3):595–607, 1994.

[30] Caroline Daly and Raymond Y Kwong. Cardiac MRI for myocardial ischemia. *Methodist DeBakey cardiovascular journal*, 2013.

[31] Christina Unterberg-Buchwald, Martin Fasshauer, Jan M Sohns, Wieland Staab, Andreas Schuster, Dirk Voit, Johannes T Kowallick, Michael Steinmetz, Jens Frahm, and Joachim Lotz. Real time cardiac MRI and its clinical usefulness in arrhythmias and wall motion abnormalities. *Journal of Cardiovascular Magnetic Resonance*, 16(Suppl 1):P34, 2014.

[32] Troy J Badger, Yaw A Adjei-Poku, and Nassir F Marrouche. MRI in cardiac electrophysiology: the emerging role of delayed-enhancement MRI in atrial fibrillation ablation. 2009.

[33] Klaus Scheffler and Stefan Lehnhardt. Principles and applications of balanced SSFP techniques. *European radiology*, 13(11):2409–2418, 2003.

[34] Uroš Vovk, Franjo Pernuš, and Boštjan Likar. A review of methods for correction of intensity inhomogeneity in MRI. *Medical Imaging, IEEE Transactions on*, 26(3):405–421, 2007.

[35] Steffen E Petersen, Paul M Matthews, Jane M Francis, Matthew D Robson, Filip Zemrak, Redha Boubertakh, Alistair A Young, Sarah Hudson, Peter Weale, Steve Garratt, et al. UK Biobank's cardiovascular magnetic resonance protocol, 2015.

[36] Steffen E Petersen, Paul M Matthews, Fabian Bamberg, David A Bluemke, Jane M Francis, Matthias G Friedrich, Paul Leeson, Eike Nagel, Sven Plein, Frank E Rademakers, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK biobank-rationale, challenges and approaches. *Journal of Cardiovascular Magnetic Resonance*, 15(1):46, 2013.

[37] Christopher M Kramer, Jorg Barkhausen, Scott D Flamm, Raymond J Kim, and Eike Nagel. Standardized cardiovascular magnetic resonance imaging (CMR) protocols, society for cardiovascular magnetic resonance: board of trustees task force on standardized protocols. *Journal of Cardiovascular Magnetic Resonance*, 10(1):35, 2008.

[38] Manuel D. Cerqueira, Neil J. Weissman, Vasken Dilsizian, Alice K. Jacobs, Sanjiv Kaul, Warren K. Laskey, Dudley J. Pennell, John a. Rumberger, Thomas J. Ryan, and Mario S. Verani. Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart. *Journal of Cardiovascular Magnetic Resonance*, 4(2):203–210, 2002.

[39] Pedro F Ferreira, Peter D Gatehouse, Raad H Mohiaddin, and David N Firmin. Cardiovascular magnetic resonance artefacts. *Journal of Cardiovascular Magnetic Resonance*, 15(1):41, 2013.

[40] Erica Dall'Armellina, Theodoros D Karamitsos, Stefan Neubauer, and Robin P Choudhury. CMR for characterization of the myocardium in acute coronary syndromes. *Nature reviews. Cardiology*, 7(11):624–636, 2010.

[41] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE signal processing magazine*, 20(3):21–36, 2003.

[42] Michael Lustig, David Donoho, and John M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[43] Tomoyuki Kido, Teruhito Kido, Masashi Nakamura, Kouki Watanabe, Michaela Schmidt, Christoph Forman, and Teruhito Mochizuki. Compressed sensing real-time cine cardiovascular magnetic resonance: accurate assessment of left ventricular function in a single-breath-hold. *Journal of Cardiovascular Magnetic Resonance*, 18(1):50, 2016.

[44] Hayit Greenspan. Super-resolution in medical imaging. *The Computer Journal*, 52(1):43–63, 2009.

[45] Michal Irani and Shmuel Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation*, 4(4):324–335, 1993.

[46] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *Image processing, IEEE Transactions on*, 13(10):1327–1344, 2004.

[47] Esben Plenge, Dirk Poot, Monique Bernsen, Gyula Kotek, Gavin Houston, Piotr Wielopolski, Louise van der Weerd, Wiro J Niessen, and Erik Meijering. Super-resolution methods in MRI: Can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? *Magnetic resonance in medicine*, 68(6):1983–1993, 2012.

[48] Dirk Poot, Vincent Van Meir, and Jan Sijbers. General and efficient super-resolution method for multi-slice MRI. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 615–622. Springer, 2010.

[49] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[50] Shantanu H Joshi, Antonio Marquina, Stanley J Osher, Ivo Dinov, John D Van Horn, and Arthur W Toga. MRI resolution enhancement using total variation regularization. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 161–164. IEEE, 2009.

[51] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Tingwen Huang. Coupled dictionary training for image super-resolution. *Image Processing, IEEE Transactions on*, 21(8):3467–3478, 2012.

[52] Jos Van Ouwerkerk. Image super-resolution survey. *Image and Vision Computing*, 24(10):1039–1052, 2006.

[53] José V Manjón, Pierrick Coupé, Antonio Buades, Vladimir Fonov, D Louis Collins, and Montserrat Robles. Non-local MRI upsampling. *Medical image analysis*, 14(6):784–792, 2010.

[54] Eric Van Reeth, Ivan WK Tham, Cher Heng Tan, and Chueh Loo Poh. Super-resolution in magnetic resonance imaging: a review. *Concepts in Magnetic Resonance Part A*, 40(6):306–325, 2012.

[55] Hayit Greenspan, Oz, Nahum Kiryati, and Sharon Peled. MRI inter-slice reconstruction using super-resolution. *Magnetic resonance imaging*, 20(5):437–446, 2002.

[56] Eyal Carmi, Siuyan Liu, Noga Alon, Amos Fiat, and Daniel Fiat. Resolution enhancement in MRI. *Magnetic resonance imaging*, 24(2):133–154, 2006.

[57] Richard Shilling, Trevor Robbie, Timothée Bailloeul, Klaus Mewes, Russell Mersereau, and Marijn Brummer. A super-resolution framework for 3D high-resolution and high-contrast imaging using 2D multislice MRI. *IEEE transactions on medical imaging*, 28(5):633–644, 2009.

[58] Richard Shilling, Senthil Ramamurthy, and Marijn Brummer. Sampling strategies for super-resolution in multi-slice MRI. pages 2240–2243, 2008.

[59] Dirk Poot, Ben Jeurissen, Yannick Bastiaensen, Jelle Veraart, Wim Van Hecke, Paul Parizel, and Jan Sijbers. Super-resolution for multislice diffusion tensor imaging. *Magnetic resonance in medicine*, 69(1):103–113, 2013.

[60] Jonghye Woo, Emi Murano, Maureen Stone, and Jerry Prince. Reconstruction of high-resolution tongue volumes from MRI. *IEEE Transactions on Biomedical Engineering*, 59(12):3511–3524, 2012.

[61] Gwendolyn Van Steenkiste, Ben Jeurissen, Jelle Veraart, Arnold J. den Dekker, Paul M. Parizel, Dirk H J Poot, and Jan Sijbers. Super-resolution reconstruction of diffusion parameters from diffusion-weighted images with different slice orientations. *Magnetic Resonance in Medicine*, 195(January 2015):181–195, 2015.

[62] Ying Bai, Xiao Han, and Jerry Prince. Super-resolution reconstruction of MR brain images. In *Proc. of 38-th Annual Conference on Information Sciences and Systems, Princeton, New Jersey*, pages 1358–1363, 2004.

[63] Matthias Ehrhardt and Marta Betcke. Multicontrast MRI reconstruction with structure-guided total variation. *SIAM Journal on Imaging Sciences*, 9(3):1084–1106, 2016.

[64] John Moore, Maria Drangova, Marcin Wierzbicki, John Barron, and Terry Peters. A high resolution dynamic heart model based on averaged MRI data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 549–555. Springer, 2003.

[65] Hsu, Cecil Yen, Ching Li, Maomin Sun, Bin Tian, and Kaygusuz. Application of wavelet-based POCS superresolution for cardiovascular MRI image enhancement. In *null*, pages 572–575. IEEE, 2004.

[66] Stefan Wesarg et al. Combining short-axis and long-axis cardiac MR images by applying a super-resolution reconstruction algorithm. In *SPIE Medical Imaging*, pages 76230I–76230I. International Society for Optics and Photonics, 2010.

[67] Stefan Wesarg and Sami Rahman. Upsampling of Cardiac MR Images: Comparison of Averaging and Super-Resolution for the Combination of Multiple Views. *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on. IEEE, 2010.*, 9:2001, 2001.

[68] Nelson Velasco Toledo, Andrea Rueda, Cristina Santa Marta, and Eduardo Romero. Super-resolution in cardiac MRI using a bayesian approach. In *SPIE Medical Imaging*, pages 866932–866932. International Society for Optics and Photonics, 2013.

[69] Oleh Dzyubachyk, Qian Tao, Dirk Poot, Hildo Lamb, Katja Zeppenfeld, Boudewijn Lelieveldt, and Rob van der Geest. Super-resolution reconstruction of late gadolinium-enhanced MRI for improved myocardial scar assessment. *Journal of Magnetic Resonance Imaging*, 42(1):160–167, 2015.

[70] Liam Cattell, Craig Meyer, Frederick Epstein, and Gustavo Rohde. Reconstructing high-resolution cardiac MR movies from under-sampled frames. In *Asilomar conference on signals, systems, and computers*, 2017.

[71] Freddy Odille, Aurélien Bustin, Bailiang Chen, Pierre-André Vuissoz, and Jacques Felblinger. Motion-corrected, super-resolution reconstruction for high-resolution 3D cardiac cine mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 435–442. Springer, 2015.

[72] Wenzhe Shi, Jose Caballero, Christian Ledig, Xiahai Zhuang, Wenjia Bai, Kanwal Bhatia, Antonio M Simoes Monteiro de Marvao, Tim Dawes, Declan O'Regan, and Daniel Rueckert. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 9–16. Springer, 2013.

[73] Wenjia Bai, Wenzhe Shi, Antonio de Marvao, Timothy JW Dawes, Declan P O'Regan, Stuart A Cook, and Daniel Rueckert. A bi-ventricular cardiac atlas built from 1000+ high resolution MR images of healthy subjects and an analysis of shape and motion. *Medical image analysis*, 26(1):133–145, 2015.

[74] Aurelien Bustin, Damien Voilliot, Anne Menini, Jacques Felblinger, Christian de Chillou, Darius Burschka, Laurent Bonnemains, and Freddy Odille. Isotropic reconstruction of MR images using 3D patch-based self-similarity learning. *IEEE Transactions on Medical Imaging*, 2018.

[75] Freddy Odille, Aurélien Bustin, Shufang Liu, Bailiang Chen, Pierre-André Vuissoz, Jacques Felblinger, and Laurent Bonnemains. Isotropic 3D cardiac cine MRI allows efficient sparse segmentation strategies based on 3D surface reconstruction. *Magnetic resonance in medicine*, 79(5):2665–2675, 2018.

[76] Ke Huang and Selin Aviyente. Sparse representation for signal classification. In *Advances in neural information processing systems*, pages 609–616, 2007.

[77] Michael Elad. *Sparse and redundant representations: From theory to applications in signal and image processing*. 2010.

[78] Kanwal K Bhatia, Anthony N Price, Wenzhe Shi, Jo V Hajnal, and Daniel Rueckert. Super-resolution reconstruction of cardiac MRI using coupled dictionary learning. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 947–950. IEEE, 2014.

[79] Spencer Brady, Minh Do, and Rohit Bhargava. Reconstructing FT-IR spectroscopic imaging data with a sparse prior. *Proceedings - International Conference on Image Processing, ICIP*, pages 829–832, 2009.

[80] Ying Song, Zhen Zhu, Yang Lu, Qiegen Liu, and Jun Zhao. Reconstruction of magnetic resonance imaging by three-dimensional dual-dictionary learning. *Magnetic Resonance in Medicine*, 71(3):1285–1298, 2014.

[81] Jose Caballero, Anthony N Price, Daniel Rueckert, and Joseph V Hajnal. Dictionary learning and time sparsity for dynamic MR data reconstruction. *IEEE transactions on medical imaging*, 33(4):979–994, 2014.

[82] Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[83] Alain Hore and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *Pattern recognition (icpr), 2010 20th international conference on*, pages 2366–2369. IEEE, 2010.

[84] Ozan Oktay, Wenjia Bai, Matthew Lee, Ricardo Guerrero, Konstantinos Kamnitsas, Jose Caballero, Antonio de Marvao, Stuart Cook, Declan O'Regan, and Daniel Rueckert. Multi-input cardiac image super-resolution using convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 246–254. Springer, 2016.

[85] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60, 2010.

[86] Darryl McClymont, Irvin Teh, Hannah J Whittington, Vicente Grau, and Jürgen E Schneider. Prospective acceleration of diffusion tensor imaging

with compressed sensing using adaptive dictionaries. *Magnetic resonance in medicine*, 2015.

[87] Jonathan Richard Shewchuk. An introduction to the conjugate gradient method without the agonizing pain, 1994.

[88] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio de Marvao, Timothy Dawes, Declan P O'Regan, et al. Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2018.

[89] Benjamin Villard, Ernesto Zacur, Erica Dal'Armellina, and Vicente Grau. Correction of slice misalignment in multi-breath-hold cardiac MRI scans. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 30–38. Springer, 2016.

[90] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[91] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[92] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

[93] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[94] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[95] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

[96] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[97] Gabriel Goh. Why momentum really works. *Distill*, 2017.

[98] Yann LeCun. Une procedure d'apprentissage pour reseau a seuil asymetrique. *proceedings of Cognitiva 85*, pages 599–604, 1985.

[99] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[100] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[101] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[102] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[103] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[104] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014.

[105] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[106] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[107] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[108] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[109] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[110] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014*, pages 184–199. Springer, 2014.

[111] Pejman Rasti, Tõnis Uiboupin, Sergio Escalera, and Gholamreza Anbarjafari. Convolutional neural network super resolution for face recognition in surveillance monitoring. In *International conference on articulated motion and deformable objects*, pages 175–184. Springer, 2016.

[112] Xiwei Huang, Yu Jiang, Xu Liu, Hang Xu, Zhi Han, Hailong Rong, Haiping Yang, Mei Yan, and Hao Yu. Machine learning based single-frame super-resolution processing for lensless blood cell counting. *Sensors*, 16(11):1836, 2016.

[113] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[114] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*, pages 391–407. Springer, 2016.

[115] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[116] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a Deep Convolutional Network for Image Super-Resolution. *Eccv*, 2014.

[117] Yang Zhao, Guoqing Li, Wenjun Xie, Wei Jia, Hai Min, and Xiaoping Liu. Gun: Gradual upsampling network for single image super-resolution. *IEEE Access*, 2018.

[118] Yifan Wang, Lijun Wang, Hongyu Wang, and Peihua Li. End-to-end image super-resolution via deep and shallow convolutional networks. *arXiv preprint arXiv:1607.07680*, 2016.

[119] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4, 2017.

[120] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*, volume 1, page 4, 2017.

[121] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Is l2 a good loss function for neural networks for image processing. *ArXiv e-prints*, 1511:8, 2015.

[122] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[123] Yapeng Tian Wei Wang Jing-Hao Xue Wenming Yang, Xuechen Zhang. Deep learning for single image super-resolution: A brief review. *arXiv preprint arXiv: arXiv:1808.03344v1*, 2018.

[124] Khizar Hayat. Super-resolution via deep learning. *arXiv preprint arXiv:1706.09077*, 2017.

[125] 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*, pages 234–241. Springer, 2015.

[126] Martin Weigert, Loic Royer, Florian Jug, and Gene Myers. Isotropic reconstruction of 3D fluorescence microscopy images using convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 126–134. Springer, 2017.

[127] Larissa Heinrich, John Bogovic, and Stephan Saalfeld. Deep learning for isotropic super-resolution from non-isotropic 3D electron microscopy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 135–143. Springer, 2017.

[128] Chi-Hieu Pham, Ronan Fablet, and François Rousseau. Multi-scale brain MRI super-resolution using deep 3D convolutional networks. 2017.

[129] Woong Bae, Seungho Lee, Gwangbeen Park, Hyunho Park, and Kyu-Hwan Jung. Residual cnn-based image super-resolution for CT slice thickness reduction using paired CT scans: Preliminary validation study. 2018.

[130] Archontis Giannakidis, Ozan Oktay, Jennifer Keegan, Veronica Spadotto, Inga Voges, Gillian Smith, Iain Pierce, Wenjia Bai, Daniel Rueckert, Sabine Ernst, et al. Super-resolution reconstruction of late gadolinium enhancement cardiovascular magnetic resonance images using a residual convolutional neural network.

[131] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[132] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[133] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

[134] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[135] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks. *arXiv preprint arXiv:1809.00219*, 2018.

[136] Jelmer Wolterink, Anna Dinkla, Mark Savenije, Peter Seevinck, Cornelis van den Berg, and Ivana Išgum. Deep MR to CT synthesis using unpaired data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 14–23. Springer, 2017.

[137] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 417–425. Springer, 2017.

[138] Sotirios Tsaftaris. Adversarial image synthesis for unpaired multi-modal cardiac data. In *Simulation and Synthesis in Medical Imaging: Second International Workshop, SASHIMI 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 10, 2017, Proceedings*, volume 10557, page 3. Springer, 2017.

[139] Ohad Shitrit and Tammy Riklin Raviv. Accelerated magnetic resonance imaging by adversarial neural network. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 30–38. Springer, 2017.

[140] Yuhua Chen, Feng Shi, Anthony G Christodoulou, Zhengwei Zhou, Yibin Xie, and Debiao Li. Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network. *arXiv preprint arXiv:1803.01417*, 2018.

[141] Maximilian Seitzer, Guang Yang, Jo Schlemper, Ozan Oktay, Tobias Würfl, Vincent Christlein, Tom Wong, Raad Mohiaddin, David Firmin, Jennifer Keegan, et al. Adversarial and perceptual refinement for compressed sensing MRI reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–240. Springer, 2018.

[142] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. GANs for medical image analysis. *arXiv preprint arXiv:1809.06222*, 2018.

[143] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial networks in medical imaging: A review. *arXiv preprint arXiv:1809.07294*, 2018.

[144] Richard Socher, Milad Mohammadi, and Rohit Mundra. Deep learning for NLP. `https://cs224d.stanford.edu/lecture_notes/LectureNotes4.pdf`, 2015. [Online; accessed 7-November-2018].

[145] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[146] Shi Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[147] Qikui Zhu, Bo Du, Baris Turkbey, Peter Choyke, and Pingkun Yan. Exploiting interslice correlation for MRI prostate image segmentation, from recursive neural networks aspect. *Complexity*, 2018, 2018.

[148] Rudra Poudel, Pablo Lamata, and Giovanni Montana. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. In *Reconstruction, Segmentation, and Analysis of Medical Images*, pages 83–94. Springer, 2016.

[149] Chen Qin, Joseph V Hajnal, Daniel Rueckert, Jo Schlemper, Jose Caballero, and Anthony N Price. Convolutional recurrent neural networks for dynamic MR image reconstruction. *IEEE transactions on medical imaging*, 2018.

[150] Dongqing Zhang, Ilknur Icke, Belma Dogdas, Sarayu Parimal, Smita Sampath, Joseph Forbes, Ansuman Bagchi, Chih-Liang Chin, and Antong Chen. A multi-level convolutional LSTM model for the segmentation of left ventricle

myocardium in infarcted porcine cine MR images. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 470–473. IEEE, 2018.

[151] Bin Kong, Yiqiang Zhan, Min Shin, Thomas Denny, and Shaoting Zhang. Recognizing end-diastole and end-systole frames via deep temporal regression network. In *International conference on medical image computing and computer-assisted intervention*, pages 264–272. Springer, 2016.

[152] Kaggle data science bowl cardiac challenge data. https://www.kaggle.com/c/second-annual-data-science-bowl/data. 2015.

[153] François Chollet et al. Keras. https://keras.io, 2015.

[154] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2, 2014.

[155] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[156] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20(1):65, 2018.

[157] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. On instabilities of deep learning in image reconstruction-does ai come at a cost? *arXiv preprint arXiv:1902.05300*, 2019.

[158] Daniel Asgeirsson, Erik Hedström, Jonas Jögi, Ulrika Pahlm, Katarina Steding-Ehrenborg, Henrik Engblom, Håkan Arheden, and Marcus Carlsson. Longitudinal shortening remains the principal component of left ventricular

pumping in patients with chronic myocardial infarction even when the absolute atrioventricular plane displacement is decreased. *BMC cardiovascular disorders*, 17(1):208, 2017.

[159] Mehran Javanmardi, Mehdi Sajjadi, Ting Liu, and Tolga Tasdizen. Unsupervised total variation loss for semi-supervised deep learning of semantic segmentation. *arXiv preprint arXiv:1605.01368*, 2016.

[160] Giacomo Tarroni, Ozan Oktay, Wenjia Bai, Andreas Schuh, Hideaki Suzuki, Jonathan Passerat-Palmbach, Ben Glocker, Antonio de Marvao, Declan O'Regan, Stuart Cook, et al. Learning-based heart coverage estimation for short-axis cine cardiac mr images. In *International Conference on Functional Imaging and Modeling of the Heart*, pages 73–82. Springer, 2017.

[161] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.