University of London
Imperial College of Science, Technology and Medicine
Department of Computing

Deep learning for Accelerated Magnetic Resonance Imaging

Krishna Gavindrajee Seegoolam

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of the University of London and
the Diploma of Imperial College, October 2022

Declaration of Originality

I declare that the work presented in this thesis is my own unless explicitly stated otherwise in the manuscript. Part of the formatting for this thesis is partially inspired from the Thesis Template by R. Robinson at `https://github.com/mlnotebook/thesis_template`.

Copyright Declaration

Abstract

Medical imaging has aided the biggest advance in the medical domain in the last century. Whilst X-ray, CT, PET and ultrasound are a form of imaging that can be useful in particular scenarios, they each have disadvantages in cost, image quality, ease-of-use and ionising radiation. MRI is a slow imaging protocol which contributes to its high cost to run. However, MRI is a very versatile imaging protocol allowing images of varying contrast to be easily generated whilst not requiring the use of ionising radiation. If MRI can be made to be more efficient and smart, the effective cost of running MRI may be more affordable and accessible. The focus of this thesis is decreasing the acquisition time involved in MRI whilst maintaining the quality of the generated images and thus diagnosis. In particular, we focus on data-driven deep learning approaches that aid in the image reconstruction process and streamline the diagnostic process. We focus on three particular aspects of MR acquisition. Firstly, we investigate the use of motion estimation in the cine reconstruction process. Motion allows us to combine an abundance of imaging data in a learnt reconstruction model allowing acquisitions to be sped up by up to 50 times in extreme scenarios. Secondly, we investigate the possibility of using under-acquired MR data to generate smart diagnoses in the form of automated text reports. In particular, we investigate the possibility of skipping the imaging reconstruction phase altogether at inference time and instead, directly seek to generate radiological text reports for diffusion-weighted brain images in an effort to streamline the diagnostic process. Finally, we investigate the use of probabilistic modelling for MRI reconstruction without the use of fully-acquired data. In particular, we note that acquiring fully-acquired reference images in MRI can be difficult and nonetheless may still contain undesired artefacts that lead to degradation of the dataset and thus the training process. In this chapter, we investigate the possibility of performing reconstruction without fully-acquired references and furthermore discuss the possibility of generating higher quality outputs than that of the fully-acquired references.

Acknowledgements

I would like to thank my supervisor Professor Daniel Rueckert and my second supervisor Professor Jo Hajnal without whom this work would not be possible. I'd also like to thank my colleagues in BioMedIA for their constant support and kindness during particularly tough times. I'd especially like to mention Dr Jo Schlemper who kept me going with late evening climbing sessions at Westway as well as providing me with inspiration during the early days of my research career.

I would also like to express my gratitude to Dr Arnaud Czaja and Dr Benoit Vanniere for introducing me to the world of scientific research as an undergraduate. Their rigorous and robust approach to conducting scientific research has stayed with me during the course of my PhD.

I cannot give thanks enough to my family and friends who were always there when I needed them. Special mentions to Tom, Callum, Margarita, Harvey, Aydan and Kasia. I should also thank my wardening team at Wilson House who provided me with some unforgettable memories and friendships during my 4-year tenure as subwarden.

Finally, I'd like to thank my partner, Hanne, for putting an everlasting smile on my face during the peaks and troughs of this PhD.

Acronyms

AUTOMAP Image reconstruction-automated transform by manifold approximation.

BiCRNN Bidirectional Convolution Neural Network.

CMR Cardiac Magnetic Resonance.
CNN Convolutional Neural Network.
CNTL Control.
CRNN Convolution Recurrent Neural Network.

DC Data Consistency.
DC-LD Data Consistent Langevin Diffusion.
DC2DDPM Data Consistent Decomposed Cascade Denoising Diffusion Probabilistic Model.
DCMAC Data Consistent Motion Augmented Cine.
DDIM Denoising Diffusion Implicit Model.
DDPM Denoising Diffusion Probabilistic Model.
DL Deep Learning.
DSC Dice Similarity Coefficient.

ED End Diastolic.
ELBO Evidence Lower Bound.
ES End Systolic.
ESC Emulated Single Coil.
ESPIRIT Eigenvalue Iterative Self-Consistent Parallel Imaging Reconstruction.

FCN Fully Connected Network.
FISTA Fast Iterative Shrinkage-Thresholding Algorithm.

GAN Generative Adversarial Network.
GMM Gaussian Mixture Model.
GPU Graphical Processing Unit.
GRAPPA Generalised Autocalibrating Partially Parallel Acquisitions.
GT Ground Truth.

HFEN High Frequency Error Norm.
HQ High Quality.
HQ-MF High Quality Motion Field.

kt-FOCUSS k-t Focal Underdetermined System Solver.

LD-DCMAC Langevin Diffused Data Consistent Motion Augmented Cine.

MAC Motion Augmented Cine.
MC Monte Carlo.
MCMC Monte Carlo Markov Chain.
ME-CNN Motion Exploiting Convolution Neural Network.

ME/MC Motion Estimation/Motion Compensation.
MF Motion Field.
ML Machine Learning.
MRI Magnetic Resonance Imaging.
MSE Mean Squared Error.
MSE-CNN Motion-Segmentation Exploiting CNN.

NN Neural Network.

PCN Parallel Coil Network.
PSNR Peak Signal to Noise Ratio.

ReLU Rectified Linear Unit.
ResNet Residual Neural Network.
RNN Recurrent Neural Network.
ROI Region of Interest.

SDE Stochastic Differential Equation.
SENSE Sensitivity Encoding.
SGD Stochastic Gradient Descent.
SN Sensitivity Network.
SSFP Steady-state free precession.
SSIM Structural Similarity Index Measure.

U-net U-net.

VB Variational Bound.
VIF Visual Information Fidelity.
VN Variational Network.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The world of medical imaging has seen recent upturn with the advent of tractable, scalable deep learning. Medical imaging is arguably one of the biggest achievements in recent human history. In the developed world, X-ray, ultrasound, magnetic resonance imaging (MRI), computed tomography (CT) and positron-emitted tomography (PET) are all commonplace. In the developing world, these technologies are slowly being introduced albeit in lower quantities and quality. Imaging is one of the most useful tools in diagnostics, particularly since they are non-invasive and in the case of non-radiation based imaging, harmless. However, the demand for diagnostics that require the use of such tools has seen a remarkable increase. In fact, the increase far outweighs the diagnostic capacity of the National Health Service (NHS) in the UK [204, 45, 55].

There are multiple studies which indicate that the main improvements to be made are in better triage and referrals to specialised diagnostic tests [191, 45]. However, a key point of the Richards' review is the need for upgraded facilities and diagnostics [166]. Furthermore, any improvements should aim to make the diagnostic process easier, more effective and more efficient.

In this thesis, we focus on improvements to the diagnostic pipeline. Diagnostics typically involves the following steps:

1. **Data Acquisition** - A specialised tool or scanner is used to acquire information about the patient's pathology

2. **Post-processing** - This typically involves using the acquired data to reconstruct a diagnostic image followed by tools to better highlight key areas of the image

3. **Analysis** - A range of tools from image segmentation to cardiac strain estimation provide useful information to a radiologist/clinician.

4. **Diagnosis** - Using all available analysis and images, a clinical radiologist produces a diagnostic report with a path to treatment

1.1 Motivation

Ultrasound and MRI are two imaging modalities that do not make use of ionising radiation. They are favourable for patients with particular needs such as pregnant women or with certain pathologies. For example, MRI may be used in cases where a CT scan cannot confirm a diagnosis of a particular cancer. Ultrasound is fast, efficient and cheap but is limited in its diagnostic capabilities. However, MRI is expensive, slow and can incur a degree of patient discomfort [186, 9].

The physics of MRI involves the use of large magnets which require a lot of maintenance and cooling. Typically, larger magnets generate images with less noise and higher quality but as a result become increasingly more expensive. The cost of such an expensive machine is subsequently - in one way, or another - passed onto the patient; the cost of the machine itself and the cost of its maintenance [112, 8, 65].

One of the things that makes MRI expensive to hospitals and to the state is not only these costs. Rather, it is the fact that only a limited number of these scans can be completed in a day due to the difficulties in acquisition. If more scans could be completed, the overall relative cost to the hospital may be reduced.

MRI acquisitions can take up to 45 minutes [89, 197, 137]. In the case of cardiac MR (CMR), this also requires electrocardiogram (ECG) gating in order to time the scanner acquisition with the phase of the heart [132]. This slow acquisition time can cause the patient a great deal

of discomfort: 1) the MRI scanner is a claustrophobic environment for many 2) keeping still for this long period of time for certain patients is very difficult 3) many types of MRI scan will require the patient to hold their breath multiple times for prolonged periods, such as with abdominal and cardiac MR - some patients are simply not able to do this. If it were possible to acquire these images more quickly, there would be a significant improvement to the cost-benefit of MR and to patient comfort.

Further to this, with the advent of point of care MRI which typically use low field strength magnets [160, 200, 202], acquisitions tend to suffer from more corruption, mostly from thermal noise by lack of a stronger, more conventional magnet [167]. In these cases, conventional MR reconstruction techniques may struggle to generate diagnostic-grade images.

MRI acquisition does not take place in the same way as a digital camera. Instead, acquisition takes place in an abstract space where individual data points represent an amalgamation of information about the resulting image. Due to the nature of the acquisition, it is possible to generate diagnostically meaningful analyses without conventionally acquiring the entirety of the data. This speed up in acquisition would allow more scans to be completed [121, 28]. This can also be combined with new MRI technologies such as Parallel Imaging (PI) where multiple parts of the data are acquired simultaneously by exploiting a type of spatial redundancy.

In this thesis, we focus on these types of acceleration in the MRI acquisition process by leveraging the data-driven approach of deep learning. We also investigate the possibility of mitigating for noisy MR scanners.

1.2 Objectives and Contributions

The objective of this thesis is to provide novel ways to perform tasks involved in the MRI process with accelerated acquisition processes. This can be achieved in a variety of different ways and hence we focus just three main concepts:

1. Exploiting motion for MRI reconstruction

2. Performing diagnostic report generation without fully sampled data

3. Probabilistic Modelling for MRI reconstruction

The thesis was undertaken mainly using data-driven approaches facilitated by deep learning algorithms. Deep learning has been the main driving force in a new-era for medical imaging.

Exploiting motion for MRI reconstruction

In cardiac MRI, it is useful to acquire a video of the heart, called a cine, rather than just a single image. However, cine acquisitions are time-consuming as we require multiple time frames of data. Furthermore, they are uncomfortable for the patient due to needing multiple breath-holds. Accelerated MRI for this case of dynamic imaging is thus of the utmost importance. In this thesis, we identify motion estimation as being a key component for increased fidelity in accelerated acquisitions. We show that cardiac motion estimation from accelerated acquisitions are good enough for functional motion-based reconstruction models to be implemented. We show that not only can we improve reconstruction quality, but also improve scan times from 45 minutes to under 1 minute (for particular use cases).

This work, spanning over Chapters 3, 4 and 6, resulted in the following publications:

- **Seegoolam, G.**, Price, A., Hajnal, J. V., Rueckert, D. (2020). Deep Learning for Robust Accelerated Dynamic MRI Reconstruction for Active Acquisition Pipelines. Abstract 1003, 29th Annual Meeting and Exhibition International Society of Magnetic Resonance in Medicine, 2020.

- **Seegoolam, G.**, Schlemper, J., Qin, C., Price, A., Hajnal, J. V., Rueckert, D. (2019). Exploiting Motion for Deep Learning Reconstruction of Extremely-Undersampled Dynamic MRI. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019.

Diagnostic Report Generation for Acceleration MRI acquisitions

MRI is used as an imaging tool for diagnostic purposes. However, smart diagnosis does not necessarily require an intermediate image reconstruction phase. The data generated in the image reconstruction is all contained within the acquisition signal. Hence, it should be possible to streamline the diagnostic process by going straight from the acquisition signal to the diagnosis. In particular, we leverage diagnostic text reports straight from multiple clinical radiologists in order to test the hypothesis. We create an image captioning model that is capable of generating accurate diagnostic text reports using only a fraction of the conventionally acquired signal.

This work, found in Chapter 5, resulted in the following publication:

- Gasimova, A. [†], **Seegoolam, G.** [†], Chen, L., Bentley, P., Rueckert, D. (2020). Spatial semantic-preserving latent space learning for accelerated DWI diagnostic report generation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020.

Probabilistic Modelling for MRI reconstruction

The question of loss functions or objectives in optimisation MRI reconstruction pipelines is one of long debate. There have been several works which introduce many different loss functions and evaluation metrics which each have certain strengths and weaknesses. A large part of this is caused by the lack of perfect, real MRI data - there are always imperfections in the deep learning training process. In this chapter, we approach the use of probabilistic modelling to mitigate for this underlying corruption and to achieve greater fidelity image reconstructions whilst also accelerating the MRI acquisition process.

This work is main contained in chapters 6 and 7. There are currently no publications from this work but there are intentions to publish two papers with the following titles:

- Motion Exploitation for Highly Undersampled Cardiac MR Cine Reconstruction using DDPMs.

[†] These authors contributed equally to this work

- Unsupervised Corruption Mitigated Accelerated MR reconstruction using Data Consistent Decomposed Cascading DDPMs.

1.3 Thesis overview

The overall thesis is structured as follows:

- Chapter 2 - A background into MRI and the tools in deep learning required to understand the contents of this thesis
- Chapter 3 - An investigation about the incorporation of motion in the reconstruction process for dynamic MRI. We investigate two types of approach using vanilla CNN and recurrent units.
- Chapter 4 - Focus on improving our proposed motion-based reconstruction algorithm by ultimately refining the motion estimate. We investigate a variety of different existing methods and also explore how abundant segmentation data can be used to influence the learning process for better quality cine reconstructions.
- Chapter 5 - Introduces the idea of accelerated MRI for diagnostic report generation. We show that we can learn from unstructured reports straight from clinical radiologists and generate a report generation model that uses only accelerated acquisitions.
- Chapter 6 - Explores the use a new class of probabilistic models called diffusion models for the cine reconstruction of accelerated dynamic MR data. We combine the ideas from Chapter 3 into the diffusion model to generate an overall more powerful reconstruction model
- Chapter 7 - We conduct a thorough investigation of diffusion models for high fidelity MRI reconstruction. A series of controlled experiments are ultimately combined to generate a powerful probabilistic model from accelerated MRI reconstruction. This is evaluated on a large knee MRI dataset which contains noise and other corruptions.

- Chapter 8 - This chapter concludes the thesis with an overview of the achievements in the thesis and possible directions for future work

Chapter 2

Background

2.1 Introduction

In this chapter, we aim to provide an overview of the problem statement and the techniques used in the thesis for the reader to understand the content in the succeeding chapters. The literature in MRI reconstruction is vast, just as in the literature in deep learning. In order to understand this thesis, the main points that will be reviewed in this introduction are:

1. **MRI data acquisition and reconstruction** — k-space data acquisition, reconstruction, parallel imaging
2. **Diffusion Models** — a method for generative modelling
3. **Image quality metrics** — PSNR, SSIM, HFEN, VIF

2.2 MRI data acquisition and reconstruction

In this section, we outline the basics of MR for image reconstruction. Data acquisition in MR does not occur in the same way as with a digital camera. Rather than imaging data being acquired pixel by pixel, MR acquires data in frequency space, similar to radio communications. This frequency space is the Fourier space or more commonly called the k-space. Further to this, data is typically acquired in shots whereby in a single shot, multiple k-space frequencies are acquired with the trajectory of the data acquisition in k-space chosen by the user. Some trajectories will be more beneficial for imaging certain anatomical parts; for example, radial trajectories are preferable for brain imaging. The functionality of the MRI scanner, such the acquisition of the k-space data and the specification of the sampling trajectories, depend on large magnets, magnetic field gradients, radio-frequency (RF) pulses and magnetisation dynamics. The Bloch equation describes the Larmor precession of magnetisation under the influence of a magnetic field in a similar way to moments/torque acting on a gyroscope. We refer the reader to [58] for a detailed overview of the physics of MR acquisitions. Some may also find [4] and [40] to be a useful reference.

In brief, different tissues types generate different signals in k-space. This is controlled by two fundamental properties of the tissue known as T_1 and T_2 which are the relaxation times for the tissue's magnetisation in the longitudinal and transverse planes. Different tissues have different relaxation times allowing them to be distinguished during the image reconstruction process. Furthermore, since different tissues have magnetisations that decay at different rates, it is possible to generate images of varying contrast by designing when to make the acquisition of the signal. For example, in T2 weighted imaging, fat and water generate large signals that appear bright in images. T1 weighted images on the other hand generate large signals for only fat. Proton density weighted images generate high signals in areas of high hydrogen density by minimising the impact of T1 and T2 decay.

Diffusion weighted images (DWI) instead depend on a more involved mechanism in order to detected areas of restricted diffusion [75]. With ischaemic brain tissue, sodium accumulates

within cells causing water to build up within the cell. Water outside of the cell is able to move freely. During DWI, a magnetic gradient pulse is applied to induce a phase shift in the precession of the magnetisation. As water moves around the body, an opposite pulse is applied to remove the phase shift applied. However, water that has moved will experience a different phase shift to the one that was applied and thus will have gain a net phase shift. The Brownian motion of water means that the phases acquired should be random and thus dephased causing the accumulated signal generated to be small or zero. However, for water that has not moved, such as that trapped in cells due to restricted diffusion, will generate large signals as these water particles will be locally in-phase.

The process of applying RF pulses and magnetic gradients followed by the signal acquisition is known as an MR sequence. Different MR sequences are more suited to weighting particular tissue properties than others and ultimately depends on the type of contrast required in the resulting image. [79, 185, 7] provides a suitable introduction into the most common MR sequences.

2.2.1 MRI Data Acquisition

The magnetic gradients, $\mathbf{G}$, of the MRI scanner dictate the k-space trajectory taken during the scan as shown in equation (2.1).

$$\mathbf{k}(t) = \frac{\gamma}{2\pi} \int_0^t G(t')dt'. \tag{2.1}$$

From this, it can be understood how the signal read by the receiver coils, $y(t)$, can instead be written as a function of k-space position instead of time, i.e. $y(\mathbf{k})$. Fundamentally, $y(\mathbf{k})$ can be related to the image/tissue magnetisation, m, via magnetisation dynamics. More specifically, the closed form solution to the Bloch equation results in a Fourier relationship between y and m. This is shown in equation (2.2).

$$y(\mathbf{k}) = \mathrm{Re}\Big[\int_{\mathbf{r}\in\mathcal{R}^3} m(\mathbf{r})e^{-i2\pi\mathbf{k}\cdot\mathbf{r}}\mathrm{d}\mathbf{r}.\Big] \tag{2.2}$$

It should be noted that typically m is complex-valued with orthogonal coils used to measure the transverse magnetisation in quadrature. This reduces signal noise and consequently generates a phase that rotates with the transverse magnetisation of the system[1]. Phase can offer useful imaging information such as in susceptibility-weighted imaging [109]. This phase is represented by converting m into a complex value $\mathbf{m}$ and y to $\mathbf{y}$. Chapter 7 of [58] provides a detailed mathematical derivation of this.

$\mathbf{y}$ is commonly referred to as the k-space. By dimensional analysis it can be seen that the space $\mathbf{k}$ is measured in units of inverse distance. The short discrete form of equation (2.2) can be written as:

$$\mathbf{y} = \mathrm{FFT}(\mathbf{m}) = \mathcal{F}\mathbf{m}. \tag{2.3}$$

Similarly:

$$\mathbf{m} = \mathrm{IFFT}(\mathbf{y}) = \mathcal{F}^H\mathbf{y}. \tag{2.4}$$

Typically, the reconstruction viewed by a radiologist is the magnitude of the image, $|\mathbf{m}|$.

2.2.2 k-space properties

Since the medical image is the inverse Fourier transform of the acquire signal, some basic Fourier properties determine the image generated. First, the field-of-view (FoV) of the image generate is determined by the spacing between data collected in k-space. The data collected in k-space occurs in three directions:

[1] when viewed in the rotating frame of reference of the MR system

- Frequency Encoded (FE) - This is a measurement of the continuous signal described above. The spacing between frequency encoded measurements, Δk_x, depends on how quickly the signal is sampled and incurs no time penalty on the overall acquisition time. In the case of Cartesian k-space sampling, this refers to the acquisition of a single line in k-space.

- Phase Encoded (PE) - This refers to the number of lines in k-space to make. The spacing between each line in k-space, Δk_y, is specified by the operator

- Slice selection - The gradient of the z-plane coil determines the thickness of the slice acquired in the subject. Stronger gradients allow more precise slices to be acquired. Specialised scanners may also use phase encoding in the z-direction but will incur significant time penalty due to the requirement of multiple slice acquisitions.

The relation between the FoV and the k-space spacing is simply:

$$\text{FoV} = m_{\max} = \frac{1}{\Delta k}. \tag{2.5}$$

Furthermore, the resolution of the image generated is determined by the maximum point in k-space acquired:

$$\text{Resolution} = \Delta m = \frac{1}{2k_{\max}}. \tag{2.6}$$

In order for the image to be reconstructed with the properties outlined above, it is necessary to introduce the Nyquist-Shannon sampling theorem: "A bandlimited continuous-time [continuous-spatial] signal can be sampled and perfectly reconstructed from its samples if the waveform is sampled over twice as fast as its highest frequency component" [142]. In terms of spatial Fourier coefficients, this is because between two sampled points, there may be higher frequency signals that could fit in between these two points. Hence, if we have knowledge of the highest frequency present, we can choose to sample at this high frequency to ensure any changes

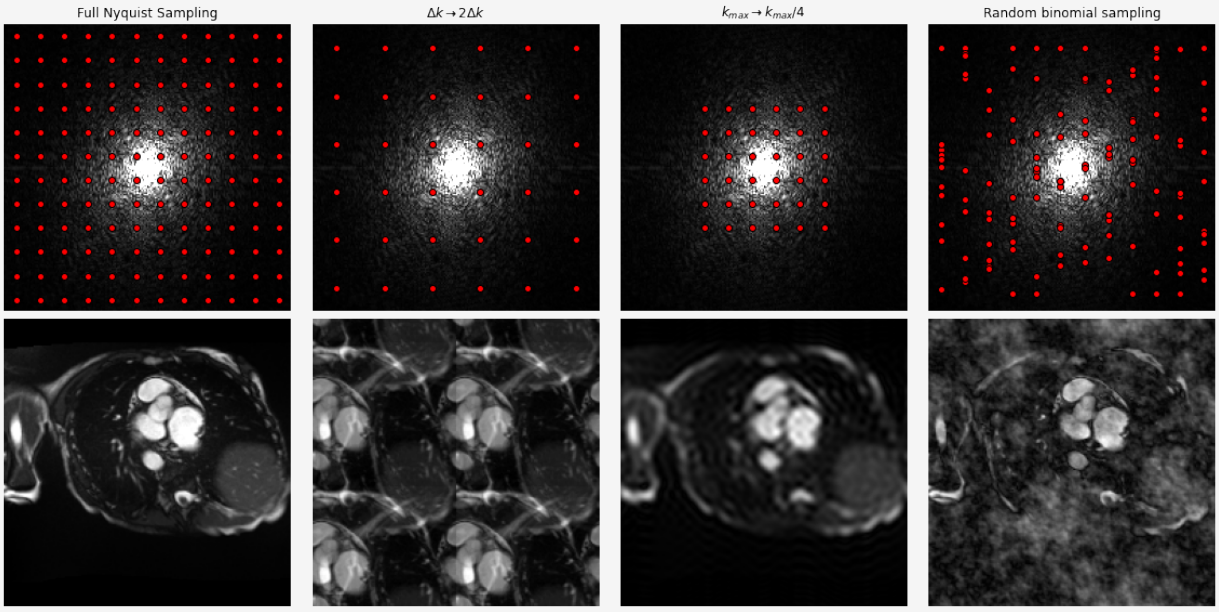

Figure 2.1: Examples of the effect of different k-space sampling on the resulting reconstruction $|\mathbf{m}|$. Top row shows the k-space and sampling points and the bottom row is the reconstruction formed from the direct IFFT followed by taking the magnitude of the complex output. The red dots indicate points where k-space samples were taken. From left to right: fully-sampled k-space with spacing Δk and maximum sample $k_{\max}$, k-spacing is sampled half as frequently, k-space is sampled only to a quarter of $k_{\max}$ but with spacing Δk, random binomial sampling. Data taken from the UK BioBank (see section 3.7.1 for more information).

between those two original points are captured. To capture all possible detail, we would need to sample at twice the frequency of the highest component since a single cycle of a waveform crosses zero at half of its time period/interval.

In terms of MRI, if the highest frequency in the scanner we sample is $k_{\max}$, then the resolution of the image will be only half as much as dictated by $k_{\max}$, as demonstrated by equation (2.6). In summary, if the signal is not properly sampled, it would violate the Nyquist-Shannon sampling criteria and thus introduces imaging artefacts into the reconstruction such as in Figure 2.1 [40, 64, 56].

2.2.3 Compressed Sensing

Whilst the Nyquist-Shannon sampling theorem requires a certain number of measurements to perfectly reconstruct the image, the underlying k-space signal is highly compressible [28, 36]. Compressed Sensing (CS) is based on the concept that the sensor signal can be sparsely represented and thus we can drastically reduce the number of sensor measurements made [31, 30]. There are three system properties required to fulfil the requirements for the CS-based image reconstruction [40, 56]:

1. *Sparsity*: The sensor signal, i.e. k-space, must have a sparse representation. Some methods may transform the sensor space into an even more sparse representation

2. *Incoherence*: The direct reconstruction of the image from undersampled k-space (called a zero-filled reconstruction) should result in aliasing artefacts that are incoherent and appear as though they are noise

3. *Nonlinear reconstruction*: Rather than use a direct IFFT, the image should be reconstructed using a non-linear optimisation algorithm that ensures the reconstruction is consistent with the acquired data in k-space but also enforces the sparsity in the transform domain.

The study in [28] is the first to apply this concept to MRI reconstruction. Rather than fully acquire the k-space required by the Nyquist-Shannon sampling theorem for the desired properties of the image reconstruction, lines in a Cartesian k-space are randomly sampled. The reconstructed image m is then obtained by solving the following optimisation problem:

$$\min ||\Psi(\mathbf{m})||_1 s.t. ||\mathcal{F}\mathbf{m} - \mathbf{y}||_2 < \epsilon. \tag{2.7}$$

Here Ψ is the sparsity transform which is commonly chosen to be a wavelet transform or finite differences and is typically solved using a Langrange multiplier λ.

2.2.4 Conventional Reconstruction Algorithms and Parallel Imaging

A more generalised form for the problem of MR reconstruction from undersampled k-space is:

$$\arg\min_{m} \lambda||\mathbf{Em} - \mathbf{y}||_2^2 + \mathcal{R}(\mathbf{m}), \tag{2.8}$$

where $\mathbf{E} = \mathbf{D}\mathcal{F}$ represents the application of the Fourier transform and the k-space undersampling mask, $\mathbf{D}$ and $\mathcal{R}$ is a regularisation term (such as the sparsity requirement). This provides us with the solution to equation (2.3) which is rewritten below:

$$\mathbf{y} = \mathbf{Em} + \delta, \tag{2.9}$$

where δ is a Gaussian noise caused by measurement imperfections (which manifests as a Rician noise in single-coil magnitude images). For clarity, $\mathbf{m}$ is a column vector of complex values that can be reshaped into the desired format (image or volume): $\mathbf{m} \in \mathbb{C}^{hwd}$, where h, w, d are the height, width and depth of the volume/image. Similarly, $\mathbf{y} \in \mathbb{C}^M$ where M is the number of k-space measurements made. In the fully-sampled case, $M = hwd$. In the oversampled case, $M > hwd$ and (2.9) can be solved using the Moore-Penrose inverse which can be found using SVD (or a range of other faster methods) [53, 38]. In the undersampled case, $M < hwd$ which is the focus in this thesis.

The choice of regularisation term $\mathcal{R}$ changes the solutions generated. An L1 norm recovers LASSO regularisation whilst an L2 norm recovers ridge regression, a special case of Tikhonov regularisation [14]. Total variation (TV) can also be used but tends to generate cartoon-like artefacts and should be considered case-by-case [39, 72, 33, 47, 6]. Wavelet transforms have also been popular in CS-MRI [48] but more recently, dictionary learning has been shown to generate high fidelity reconstruction, particularly in the case of spatio-temporal reconstructions [73, 54].

In terms of completing the above optimisation, a variety of different methods have been ex-

plored. Naive gradient descent requires convexity and differentiability which are not always satisfied by choice of regularisation e.g. L1 norm. The (fast) iterative shrinkage-thresholding algorithm ([F]ISTA) [41] is a well-studied approach to the optimisation in the case of L1 regularisation by use of proximal gradient descent and a shrinkage operator. Many other ISTA-based methods have since been proposed such as backtracking (B)ISTA, eigenvalue-free (EF)ISTA and fast (F)EFISTA [41, 195].

Alternating directions method of multipliers (ADMM) is another popular method for optimisation whereby an auxiliary variable introduces convex relaxation which allows the subsequent problem to be solved with the augmented Langrange method. Unlike ISTA, the choice of regularisation is far ranging from TV to dictionary learning [73]. Variable splitting approaches also involve introducing auxiliary variables for convex relaxation resulting in multiple subproblems that can be solved individually [44, 199, 103, 123].

Parallel Imaging is a paradigm in which multiple receiver coils are employed around the scanner to capture data in a locally-sensitive manner. Each coil is accompanied by an image-domain sensitivity map which $\mathbf{m}$ is subject to. This results in extra information available to the reconstruction process where coil sensitivities embedded in the acquired data provides extra redundant spatial information. In general, the forward model is written as $\mathbf{E} = \mathbf{U}\mathcal{F}\mathbf{S}\mathbf{m}$ where S represents the coil sensitivities in the spatial domain. This noise reduction is not so high if the data is undersampled. The signal-to-noise ratio (SNR) in the reconstructed image using parallel imaging is:

$$\mathrm{SNR}_R = \frac{1}{g\sqrt{R}}\mathrm{SNR}_1, \tag{2.10}$$

where R is the rate at which data is undersampled, g is a geometry dependent noise amplification factor and SNR_1 is the SNR of the fully-sampled acquisition.

In the case of compressed sensing parallel imaging, various methods have been explored but mostly fall into the category of SENSE-based methods where coil sensitivity maps are estimated for the reconstruction or GRAPPA-based methods where missing k-space values are estimated

using a specified kernel defined by a set of autocalibrating lines (ACL) in k-space. ESPIRIT is a method that combines the GRAPPA and SENSE using an eigenvalue approach to solving for the coil sensitivity maps explicitly [21, 16, 62].

2.2.5 Dynamic Reconstruction

Whilst the above reconstruction techniques can be used in the case of cine MRI, where one wishes to reconstruct a sequence of images to form a video, there are specific reconstruction algorithms that exist.

k-t BLAST and k-t SENSE reconstruct dynamic sequences by exploiting spatiotemporal correlations in m-f space (spatial image-temporal frequency space) which in the literature is referred to as x-f space. In particular, a filter is devised using a calibration signal (low resolution reconstruction) along with knowledge of the signal covariance to estimate how signals in k-t space are distributed in x-f space [24, 131]. This is subsequently generalised to k-t FOCUSS where the signal covariance term is replaced with an iteratively updated weighting matrix [35, 42]. k-t SLR is another method which exploits x-f space sparsity but combines it with a spectral decomposition prior [51]. Dictionary learning has also been combined with temporal gradient sparsity for dynamic CS MRI reconstruction [54].

2.3 Image analysis tasks post-acquisition

Once an MR image is acquired, it can be subject to numerous types of analysis. For example, segmentation of MRI brain images is vital for image-guided interventions and surgical planning [63]. The complex structure of the brain also makes segmentation an important tool for aiding in pathology diagnosis and analysis the development of the brain. In the case of cardiac imaging, quantifying the volumes of the ventricles and the ejection of blood is only possible with accurate segmentations [52, 152]. Post image reconstruction, highly quality segmentations can only be generated in the reconstruction is of high fidelity and faithful to the patient. The resulting

images may also be used for direct pathological classification without any prior segmentation. The use case of reconstruction MR images are evidently numerous and with a radiologist in between the reconstruction phase and analysis phase, a high level of interpretability is possible. Clinical reports that are written by a radiologist often summarise the findings of any analysis conducted on the image. Recently, research has been conducted to attempt to leverage the abundance of clinical report data for diagnosis by automating the radiological report generation process [130, 203, 124] which we further investigate in Chapter 5.

2.4 Diffusion Models

In this section, we introduce diffusion models as an approach for generative modelling. For a review of deep learning, we refer the reader to [179], [155] and [77].

Discriminative regressive models create implicit decision boundaries in the data space allowing data points to be discriminated. Typically, these models generate the value which is of the greatest likelihood rather than producing a distribution of values. For example, an L2 training loss - often used in MRI reconstruction training - is directly related to maximum likelihood estimation (MLE) under the assumption of normally distributed errors as shown in equation (2.11).

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X},\theta) &= \prod_i p(\mathbf{y}_i|\mathbf{x}_i,\theta) \\ &= \prod_i p(\mathbf{y}_i - \mathbf{f}_\theta(\mathbf{x}_i)) \\ &= \prod_i \mathcal{N}(\mathbf{y}_i - \mathbf{f}_\theta(\mathbf{x}_i), \sigma^2) \\ \log p(\mathbf{Y}|\mathbf{X},\theta) &= C - \frac{1}{2\sigma^2}\sum_i ||\mathbf{y}_i - \mathbf{f}_\theta(\mathbf{x}_i)||^2, \end{aligned} \tag{2.11}$$

where $\mathbf{x}$ is the random variable that conditions the model, $\mathbf{y}$ is the random variable we wish to predict, θ are learned parameters/weights of the predictor f, σ is the noise level in the data

and C is a constant. Hence, an MRI reconstruction model trained with an L2 loss will generate a singular output that has maximised the likelihood. (The key assumption here is identically, independently distributed (i.i.d.) noise for each data point which may not always be the case).

With generative models, the aim is to learn the joint distribution between the input and output, $p(\mathbf{x}, \mathbf{y})$ which is usually the data distribution, $p(\mathcal{D})$. Rather than simply discriminate between data points, generative models are able to generate new data points using the learned data distribution. Diffusion models are an approach to generative modelling whereby the score function is estimated by a NN and the score function is all that is needed in the guidance of a predefined stochastic process for data generation. This contrasts to GANs where adversarial training is employed and is typically prone to mode-collapse or other difficulties in training [183, 162, 178, 192, 171, 108]. It also contrasts to normalising flows where a combination of low dimensional latent variables, expensive training the requirement of constrained, invertible transformations mean that expressiveness is limited [164, 193]. Diffusion models are high dimensional, straightforward to train and expressive achieving high fidelity and log-likelihoods [170].

We redefine $\mathbf{x}$ as the target image to be generated for this section only. Diffusion models decompose the image generation process, $p(\mathbf{x})$, into a series of T intermediates steps whereby a latent variable $\mathbf{x}_t$ is generated according to a distribution p_t at training time or q_t at inference. The distribution $q_t(\mathbf{x}_t|\mathbf{x})$ is well-defined beforehand and the process of generating this latent variable given $\mathbf{x}$ can be summarised in the form of a stochastic differential equation:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w} \tag{2.12}$$

Here $\mathbf{x}$ is our latent variable being diffused, $\mathbf{w}$ is a Wiener process, t is the index of said process i.e. the intermediate step, f is the drift term and g is the volatility term. For certain choices of f and g, $p_t(\mathbf{x}_t|\mathbf{x})$ has a closed form expression and Fokker-Plank can be used to show that the stochastic process transforms our data at $t = 0$, $\mathbf{x}_0 = \mathbf{x}$, to something close to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [170]. This standard normal distribution forms our prior for the sampling process where a noise removal occurs at each step that gradually transforms the random noise $\mathbf{x}_T$ to our data point

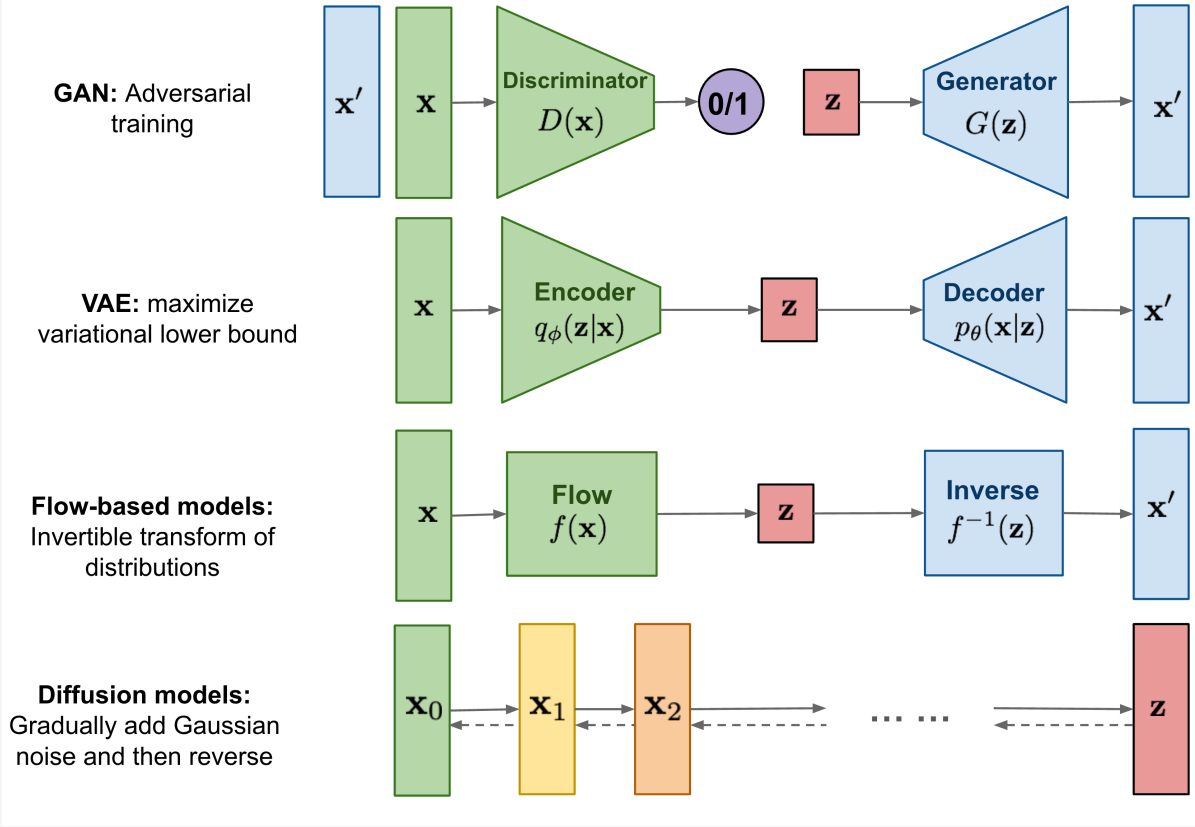

Figure 2.2: Image taken courtesy of [196]. Here, $\mathbf{z} = \mathbf{x}_T$.

$\mathbf{x}_0$. This is summarised in Figure 2.2 with an illustration in Figure 2.3.

Denoising diffusion probabilistic models (DDPMs) are a variant of this stochastic modelling process and is the flavour of diffusion model used in this thesis. The intermediate latent variables are set by predefined distributions. The forward diffusion process has a prescribed distribution as follows:

$$q_t(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{2.13}$$

Some useful variables to define the variance schedule of the intermediate distributions are: $\alpha_t = 1 - \beta_t$, $\tilde{\alpha}_t = \prod_t(\alpha_t)$ and $\beta_t = \beta_0 + \frac{(\beta_T - \beta_0)}{T}t$ where T is the chosen number of diffusion steps in the model and $\beta_0, \beta_T \ll 1$ controls the prescribed variance schedule.

It can be shown that equation (2.13) is equivalent to:

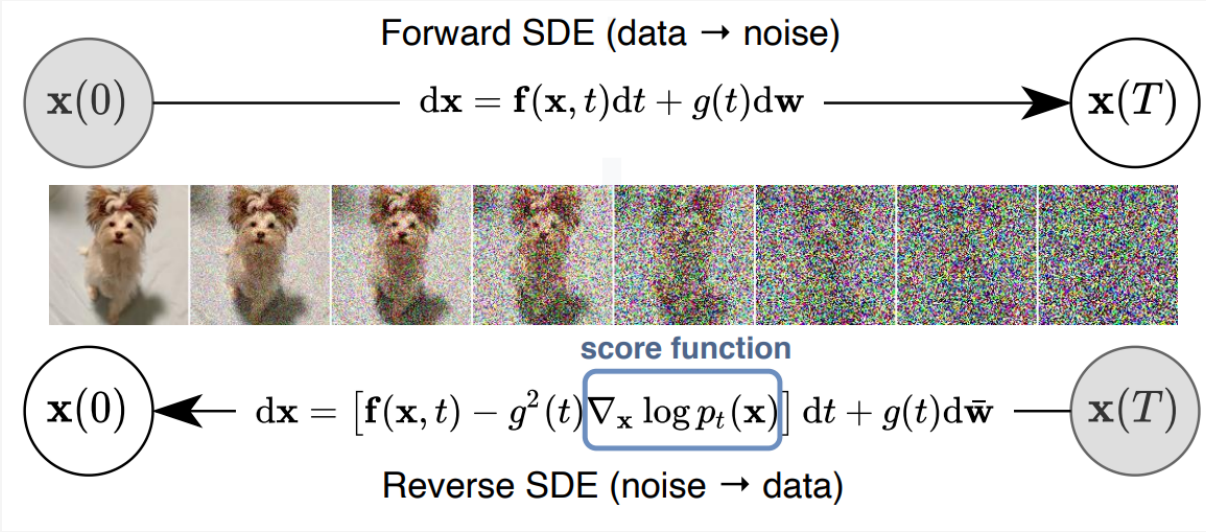

Figure 2.3: Image taken courtesy of [170]. Illustration depicting the forward and reverse diffusion process.

$$q_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\tilde{\alpha}}\mathbf{x}_0, (1-\tilde{\alpha})\mathbf{I}) \tag{2.14}$$

which means that $\mathbf{x}_t$ can be computed using the reparameterisation trick:

$$\mathbf{x}_t(\mathbf{x}_0, \epsilon) = \sqrt{\tilde{\alpha}}\mathbf{x}_0 + \sqrt{1-\tilde{\alpha}}\epsilon, \tag{2.15}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Similarly, for the reverse process:

$$p_t(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \beta_t), \tag{2.16}$$

where μ_θ is a function that is learnt via the following training process and θ are the weights of the NN. The training objective is:

$$\mathbb{E}\left[-\log p_t\left(\mathbf{x}_0\right)\right] \leq \mathbb{E}_q\left[-\log \frac{p_t\left(\mathbf{x}_{0:T}\right)}{q_t\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\right] = \mathbb{E}_q\left[-\log p_T\left(\mathbf{x}_T\right) - \sum_{t \geq 1} \log \frac{p_t\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q_t\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)}\right], \tag{2.17}$$

where we note that p_T is a standard normal distribution. This can be simplified to a sum of KL divergences between the reverse and forward distributions:

$$L = C + \sum_t L_t = C + \sum_t D_{\mathrm{KL}}(q_t(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{x}_0)||p_t(\mathbf{x}_t|\mathbf{x}_{t+1})) \tag{2.18}$$

where C is some constant.

A convenient form for $\mu_\theta(\mathbf{x}_t, t)$ is:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\tilde{\alpha}}}\Big(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\tilde{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\Big), \tag{2.19}$$

where ϵ_θ now performs the action of the NN. Combining this with equations (2.13)- (2.16) means that we can reduce our loss function to equation (2.20).

$$L = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\Big[\gamma_t||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||^2\Big] \tag{2.20}$$

where the weighting in our variational lower bound γ_t is discarded as suggested in [161] in favour of sample quality rather than log-likelihood. It has previously been suggested to keep the weighting and instead using importance sampling to reduce the variance of the bound at training time [193].

In essence, we train a neural network to denoise noisy images in a similar fashion to a denoising autoencoder. At inference, we sample from a normal distribution and use our NN to predict the noise present in the noisy image. This noise prediction is partially removed from the image at iteration $t = T$ to advance towards the next iteration $t - 1$. This repeats until $t = 0$. This noise removal is given by:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\tilde{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\Big) + \sqrt{\beta_t}\mathbf{z}, \tag{2.21}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

In the forward SDE, equation 2.12, we state the forward diffusion process transforms our data into a standard normal distribution. In the case of DDPMs, a flavour of diffusion model used in this thesis, the drift and volatility terms are set by the functions in equations (2.22) and (2.23) [170].

$$\mathbf{f}(\mathbf{x}, t) = -\frac{1}{2}\beta_t \mathbf{x} \tag{2.22}$$

$$g(t)^2 = \beta_t \tag{2.23}$$

Every forward diffusion process has a reverse process which itself is a diffusion process [1]:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_x \log q_t(\mathbf{x})]d\tilde{t} + g(t)d\tilde{\mathbf{w}}, \tag{2.24}$$

where $\tilde{w}$ and $\tilde{t}$ are in the reverse time direction. This reverse diffusion process requires the gradient of data density, i.e. the score function, to be known or learnt: $s_\theta = \nabla_x \log q_t(\mathbf{x})$. The SDE formulation in the reverse direction is equivalent to the iterative denoising autoencoder formulation above and in fact $s_\theta(\mathbf{x}_t, t) \approx -\epsilon_\theta(\mathbf{x}_t, t)/\sqrt{1-\bar{\alpha}_t}$. Using Euler-Maruyama to discretise the reverse diffusion process in equation (2.24) recovers the noise removal process in equation (2.21) showing the equivalence of score-based generative modelling and DDPMs [170, 193, 196].

2.4.1 Deep Learning Reconstruction

One typical approach to CS MRI with deep learning is replacing terms in an algorithm with a more general image denoiser. This image denoiser can take many forms such as an explicit filtering in some transform domain [34]. However, recently, deep learning based image denoisers have found success for image restoration tasks [118, 100, 115, 90]. It has been shown that

denoising CNNs (DnCNN) can be used for the task of MRI reconstruction by training them to remove noise from knee images. This is then used to replace the proximal mapping step in an ADMM-based reconstruction [149, 100]. These type of approach is termed 'plug and play (PnP)'.

Beyond the PnP approach, end-to-end CNN CS-based optimisation has been extensively studied. The conventional reconstruction algorithms discussed in section 2.2.4 have been extended with deep learning. Typically, certain steps of the optimisation are replaced with trainable CNNs. For example, direct proximal gradient descent has a variant called PGD-Net [153], variable splitting has a variant called deep-cascade CNN (DC-CNN) [93], ADMM-based optimisation has ADMM-Net [82] and FISTA has FISTA-Net [198, 154].

For parallel imaging, GRAPPA-Net was developed as an end-to-end deep learning variant of GRAPPA. Variational network (VN) is a method based on gradient descent whereby a Field-of-Experts model is generalised to form a series of convolutions with trainable activation functions [29, 110]. In contrast, Model Based Deep Learning (MoDL) network [103] was derived using Taylor-expansion based unrolling (similar to DC-CNN) but used conjugate gradient descent (CGD) to compute the multi-coil data consistency term rather than, for example, an implicitly learned proximal mapping [103, 123, 189]. Although it requires using CGD, MoDL can still be trained end-to-end (E2E). VN and MoDL both require precomputed coil sensitivity maps that can be obtained with ESPIRIT. E2E-VN [172] extended the VN to learn to generate the sensitivity maps and thus at test-time does not require any pre-computation.

Also in parallel imaging, sensitivity coil networks (SCNs) reconstruct the image directly by using precomputed sensitivity maps to enforce data consistency before recombination of individual data-consistent coil images [126]. The data consistency with SCNs can be enforced using gradient descent (like in VNs), proximal mapping (like in MoDL) or with variable splitting [123]. Parallel coil networks (PCNs) [126, 138, 176] on the other hand enforce data consistency on the coils individually analogous to DC-CNN. They reconstruct all coil images explicitly leaving the network to learn how to weight the individual coil data in the reconstruction process rather than explicitly using conjugate gradient descent and sensitivity maps to perform data

consistency [103]. The study in [151] extend PCNs by applying regularisation on an implicitly-learned coil-combined image rather than on individual coil images. For a more detailed review of parallel MRI reconstruction algorithms, we refer the reader to [163].

In this thesis, there is a focus on motion-based accelerated dynamic MRI cine reconstruction with deep learning which we are the first to approach in an E2E framework. Post-publication of the content in Chapter 3, we have been made aware of another approach to motion-based reconstruction. The work by [190] exploits motion information to warp fully-sampled reference acquisitions to aid the reconstruction process. This approach is very similar to motion compensation/correction mechanism presented in kt-FOCUSS [42]. However, it is different to the content in Chapter 3 since we do not require fully sampled references and we introduce motion as part of unrolled reconstruction optimisation itself rather than only as an independent constraint.

2.5 Image Quality Metrics

In order to evaluate the quality of the generated images, we require some quantitative metrics that can aid in making objective judgements about proposed methods. Standard metrics include the root normalise mean squared error (RNMSE) which can be calculated as follows:

$$RNMSE = \frac{1}{\sigma}\sqrt{\frac{\sum_{i,j}^{M,N}(y_{ij} - \hat{y}_{ij})^2}{MN}}, \tag{2.25}$$

where y_i is the reference image/ground truth, $\hat{y}$ is the prediction, MN are the total number of pixels in the image and $\sigma = y_{max} - y_{min}$ is the data range.

Another popular metric is the PSNR, which for data normalised between 0 and 1 is given as:

$$PSNR = -10\log_{10}\Big[\frac{1}{MN}\sum_{i,j}^{M,N}(y_{ij} - \hat{y}_{ij})^2\Big]. \tag{2.26}$$

Whilst these metrics are easy to interpret, [25] proposes to develop metrics that are closer in line with the human visual system (HVS). The image is decomposed into a series of wavelet coefficients $C^{n,j}$ where j is the subband and n is decomposition number from a Gaussian Scale Mixture Model (GSM; see [20]). Combining this with a distortion model allows the information present in the reference and image prediction to be calculated. The visual information fidelity (VIF) metric is a ratio between the information extracted from the test image and the reference image [25][2]. In particular, it should be noted that VIF was developed for use with videos.

Further work into the HVS led to the development of another popular metric known as the structural similarity index metric (SSIM) [26]:

$$SSIM(x,y) = l(x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma}, \tag{2.27}$$

where luminance, contrast and structure expressions are given below:

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{2.28}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{2.29}$$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}, \tag{2.30}$$

where c_i are constants, and μ and σ are the pixel mean and standard deviation/covariance in a chosen window size W. In this thesis, the SSIM is calculated with $\alpha = \beta = \gamma = 1$ and we report the average SSIM across all sliding windows of the image. We use the SSIM implementation provided in the Python `skimage` package.

Finally, in [49], the high frequency error norm (HFEN) was used to quantify the quality of

[2]The implementation used in this thesis can be found at `https://github.com/aizvorski/video-quality/blob/master/vifp.py`

edge information and small details. This involves convolving the Laplacian of Gaussians (LoG) filter with the images to highlight high frequency information. HFEN is then the L2-norm of difference between LoG filtering the reference image and prediction[3]:

$$HFEN = \frac{\sum_{ij} ||[\text{LoG}(y) - \text{LoG}(\hat{y})]_{ij}||^2}{\sum_{ij} ||[\text{LoG}(y)]_{ij}||^2}. \tag{2.31}$$

[3]The implementation for LoG used in this thesis can be found at `https://github.com/styler00dollar/pytorch-loss-functions/blob/main/vic/filters.py`

Chapter 3

ME-CNN: Motion Exploiting Convolution Neural Networks for Motion-based accelerated MR cine reconstruction

The problem of accelerated acquisition for cine MRI has been recently tackled with deep learning techniques. However, current state-of-the-art approaches do not incorporate a strategy to exploit the full temporal information of cine MRI which can aid in producing higher quality cine reconstructions. In this paper, we propose a novel method for exploiting the full temporal dynamics of cine MRI for reconstruction. Specifically, motion estimations are derived from undersampled MRI sequences. These are used to fuse data along the entire temporal axis to produce a novel data-consistent motion-augmented cine. This is generated and utilised within an end-to-end trainable deep learning framework for MRI reconstruction. We also explore a recurrent network approach to this problem with promising results.

3.1 Introduction

The problem of image reconstruction for accelerated MRI has been a well-explored problem with approaches ranging from sensitivity encoding to Bayesian dictionary learning [16, 54]. The process of MR image reconstruction typically involves the reconstruction of an image, vectorised as $\mathbf{m}$, from a set of data collected in the k-space of the scanner, $\mathbf{y}$. The physics of MR imaging relates $\mathbf{m}$ and $\mathbf{y}$ through an encoding matrix $\mathbf{E}$ which applies coil-sensitivity maps, k-space sampling mask and a Fourier transform to the image domain. This can be summarised as:

$$\mathbf{y} = \mathbf{E}\mathbf{m} + \epsilon, \tag{3.1}$$

Here ϵ represents the noise in the data acquisition. Since MRI data acquisition takes place in k-space, accelerated acquisition typically involves acquiring fewer samples in k-space whilst trying to reconstruct with the same resolution in operator-specified image space. This is known as undersampling. Given the acquired undersampled k-space data, $\mathbf{y}$, the violation of the Nyquist sampling criterion means that the process of finding the true image, $\mathbf{m}$, is thus ill-posed and typically ill-conditioned[1]. In order to estimate the underlying image, $\mathbf{m}$, typically a regularisation term is added in order to guide the optimisation process to a set of plausible reconstructions:

$$\frac{\alpha}{2}\sum_{i}||\mathbf{E}\mathbf{m} - \mathbf{y}||_2^2 + \mathcal{R}(\mathbf{m}), \tag{3.2}$$

where α is a hyperparameter to control the balance between the data term and the regularisation term.

In this chapter, we explore the use of motion estimation to better exploit the temporal direction of dynamic MR cine acquisitions. In particular, we do not require intricate motion estimates from sophisticated methods such as MR tagging ([5]) but instead derive motion estimates

[1]The least-squares solution with the Moore-Penrose pseudoinverse is extremely sensitive to perturbations in the acquired data e.g. scanner noise

directly from the undersampled acquisition. We modify the optimisation in equation (3.2) in order to incorporate motion into the reconstruction process.

3.2 Related work

More recently, there has been a shift towards deep learning for the reconstruction of MRI via compressed sensing approaches. Since MRI data acquisition takes place in Fourier space, also known as 'k-space', this generally involves acquiring fewer samples in k-space whilst trying to reconstruct with the same resolution in image space. However, a direct zero-filled reconstruction leaves behind aliasing artefacts which drastically reduces perceptual quality. Instead, deep learning has been used to recover useful information from aliasing artefacts and subsequently improve image quality [93, 110, 113]. For dynamic MRI, the current widely accepted state-of-the-art is the DC-CNN studied by Schlemper *et al.* (2018) which uses cascades of convolutional neural networks, with a residual connection from the input of each cascade to its reconstruction output. In addition to this, a *data consistency* (DC) step is applied to ensure the output of each cascade is consistent with the original k-space information [93].

In order to exploit the entire temporal domain, motion field estimation is required as will be explained in section 3.3. Motion estimation has been used in several studies for the purpose of correcting motion corrupted acquisitions [27, 32, 78]. A framework called k-t FOCUSS introduce a way to use motion estimation for the purpose of cine MRI reconstruction however it requires fully-sampled reference frames (see section 4.1.3 for more details; [42, 35]). Qin *et al.* (2018) studied the use of unsupervised learning for motion estimation in order to perform joint cardiac MRI segmentation and motion estimation [114, 74]. This study used fully sampled MRI cines within a VGG architecture trained to produce an optical flow estimate between a given frame and a target. The motion estimation technique used was based on Ahmadi *et al.* [74]. In this study, we combine the methodology by Ahmadi *et al.* (2016) and Schlemper *et al.* (2018) to a produce a novel, end-to-end-trainable *motion-estimating convolutional neural network*, or ME-CNN, which can reconstruct extremely undersampled MRI cines. By explicitly exploiting motion, we contribute towards building MRI reconstruction models that can harness

the full temporal domain of the original k-space sequence.

3.3 Unrolled Motion-based Optimisation

Typically, the optimisation process for cine MR reconstruction can be posed as the same inverse problem from section 3.1. However, there is no explicit mechanism to relate one temporal frame to another. We introduce a regularising condition binding the k-space of neighbouring frames allowing the optimisation to be written as in equation (3.3). In this scenario, we consider the case of single coil imaging with undersampling mask, D, such that encoding matrix $E = D\mathcal{F}$.

$$\frac{\alpha}{2}\sum_i ||D\mathcal{F}m_i - y_i||_2^2 + \frac{\beta}{2}\sum_i ||D\mathcal{F}M_i m_i - y_{i+1}||_2^2 + \mathcal{R}(M, m), \tag{3.3}$$

where i denotes the time frame index, β is a hyperparameter and M_i is the motion matrix that warps an image-space frame from time frame i to $i+1$. We note the relationship between frames in equation (3.4).

$$M_i m_i = m_{i+1} \forall i \in \{1 \dots T\}, \tag{3.4}$$

where T is the number of temporal frames in the acquisition.

In order to better understand possible approaches to this optimisation problem, we deconstruct it into multiple sub-problems using a Lagrange multiplier to link the sub-problems together. This type of decomposition is sometimes known as variable splitting or Langrangian Decomposition [2, 3, 123]. Auxiliary variables u_i and x_{i+1} are introduced which allows an explicit denoising problem to be formulated separate from any other steps such as data consistency. The bounds on these new auxiliary variables are such that equations (3.5) and (3.6) are enforced.

$$u_i = m_i \tag{3.5}$$

$$x_{i+1} = M_i m_i = M_i u_i \tag{3.6}$$

Given the conditions of equations (3.5) and (3.6), the initial optimisation can be written as equation (3.7).

$$\frac{\alpha}{2}\sum_i ||D\mathcal{F}m_i - y_i||_2^2 + \frac{\beta}{2}\sum_i ||D\mathcal{F}x_i - y_i||_2^2 + \mathcal{R}(M) + \mathcal{R}(u), \tag{3.7}$$

where the motion and image regularisation have been decoupled into separate terms.

Using a simple convex L2 regularisation to for equations (3.5) and (3.6), the full optimisation, O, becomes as shown in equation (3.8).

$$\begin{aligned} O(m, M; y) =& \frac{\alpha}{2}\sum_i ||D\mathcal{F}m_i - y_i||_2^2 \\ &+ \frac{\beta}{2}\sum_i ||D\mathcal{F}x_i - y_i||_2^2 + \frac{\gamma}{2}\sum_i ||u_i - m_i||^2 \\ &+ \frac{\sigma}{2}\sum_i ||M_i m_i - x_{i+1}||^2 + \mathcal{R}(M) + \mathcal{R}(u), \end{aligned} \tag{3.8}$$

where γ and σ are hyperparameters. γ, α, σ, and β are terms which control the faithfulness of the reconstruction to the acquired data and are commonly referred to as the noise terms. In the case of noiseless data, β and α approach ∞ and σ and γ can be absorbed into the regularisation terms.

Four sub-problems can be formed from (3.8), shown in equations (3.9)-(3.12).

$$\textbf{image denoising} \qquad u = \arg\min_{u} \left\{ \frac{\gamma}{2} \sum_{i} ||u_i - m_i||^2 + \frac{\sigma}{2} \sum_{i} ||M_i m_i - x_{i+1}||^2 + R(u) \right\} \tag{3.9}$$

$$\textbf{motion estimation} \qquad M = \arg\min_{M} \frac{\sigma}{2} \sum_{i} ||M_i m_i - x_{i+1}||^2 + R(M) \tag{3.10}$$

$$\textbf{data consistency} \qquad m = \arg\min_{m} \frac{\beta}{2} \sum_{i} ||D\mathcal{F} m_i - y_i||^2 + \frac{\gamma}{2} \sum_{i} ||u_i - m_i||^2 \tag{3.11}$$

$$\textbf{motion-reconstruction} \qquad x = \arg\min_{x} \left\{ \frac{\alpha}{2} \sum_{i} ||D\mathcal{F} x_{i+1} - y_{i+1}||^2 + \frac{\sigma}{2} \sum_{i} ||M_i u_i - x_{i+1}||^2 \right\} \tag{3.12}$$

Solving these equations lead to closed-form solutions for equations (3.11) and (3.12). Equation (3.9) is replaced with a CNN denoiser and (3.10) is replaced with a CNN motion estimator. The solution for (3.11) is a denoising reconstruction following by the requirement that the k-space of the reconstruction, u_i^k is consistent with the acquired data at this frame, y_i. This is the closed form solution in equation (3.13). This is also known as data consistency/fidelity as recovered in prior studies [93, 49, 177].

$$m_i^k = (\alpha \mathcal{F}^T D^T D \mathcal{F} + \sigma I)^{-1} (\alpha \mathcal{F}^T D^T y_i + \sigma u_i^k) \tag{3.13}$$

The solution for (3.12) applies the motion estimate to a reconstructed frame to produce the next

frame in the temporal sequence. It then applies data consistency with the acquired data at this frame, y_{i+1}. This is a closed-form solution, equation (3.14). We refer to this step and variants of this as the 'DCMAC' step which is an abbreviation for data-consistent motion augmented cine.

$$x_i^k = (\alpha \mathcal{F}^T D^T D \mathcal{F} + \sigma I)^{-1} (\alpha \mathcal{F}^T D^T y_{i+1} + \sigma M_i^k u_i^{k-1}) \tag{3.14}$$

In other words, equation (3.14) represents taking the output from the previous cascade/iteration, the denoised frame, and warping it to the next frame where data consistency is subsequently applied. The output from the previous cascade/iteration can be used due to the enforcement of the condition $m_i = u_i$ in equation (3.5).

Finally, equations (3.5) and (3.4) allows the expression $M_{i-1}u_{i-1} \approx u_i$ to hold meaning that sub-problem in equation (3.9) can be off-loaded to a CNN denoiser with parameters θ. The CNN takes as input: the output of the previous cascade, $k-1$, and the output of the DCMAC operation in equation (3.14). These are used to generate a new estimate for the reconstruction as shown in equation (3.15). This then followed by the data consistency step above in equation (3.13) before the whole process repeats in an iterative motion-based reconstruction process. Figure 3.1 illustrates this process.

$$u^{k+1} = u^k + \text{CNN}_\theta(u^k, x^k) \tag{3.15}$$

3.4 Deep Learning implementation for exploiting temporal consistency

The decomposition in section 3.3 exploits the temporal dimension for image/cine reconstruction. We would require at least T CNN blocks to ensure that at least every frame is exploited at least once. Typical cines contain $T > 30$ frames which would mean fitting 30 or more CNN blocks

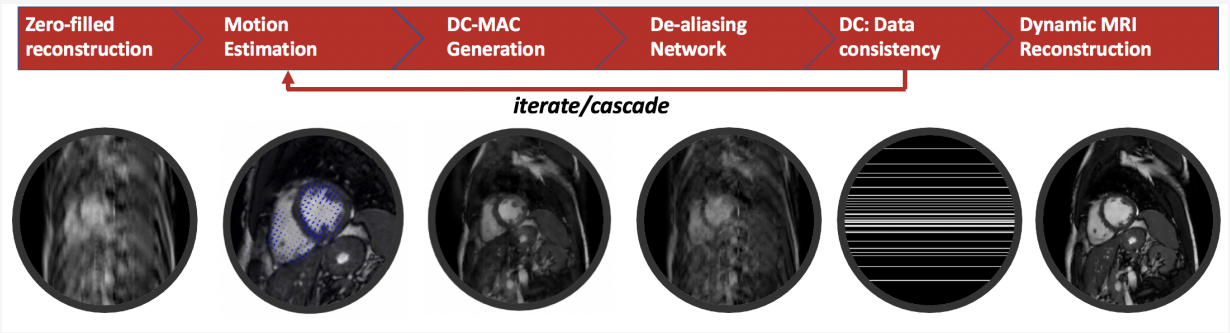

Figure 3.1: The unrolled motion-based reconstruction architecture. Each CNN contains a motion estimation block, a DC-MAC generator, a set of dealiasing/denoising convolution layers and a data consistency layer. More information can be found in section 3.5.

into GPU memory. In our studies in this chapter, we use vanilla CNN layers that consume relatively little memory compared with the U-net architecture. In spite of this, fitting into GPU memory this many CNN cascades is not possible.

One possible way to ensure that we are able to better exploit every frame in our data acquisition, is to reduce the number of CNN blocks used. More specifically, we can skip the CNN denoiser block N_{dcmac} times before using the next CNN denoiser block[2]. The total number of iterations would be $N_{\text{dcmac}} \times N_c$ where N_c is the number of CNN blocks (cascades) used. In our initial study, we focus on extreme acceleration, $A_f = 51$ for $N_c = 5$ before investigating further with $N_c = \{3, 5\}$ for a large range of acceleration rates $A_f = \{4...51\}$.

The interpretation of this would be to apply the closed form equation (3.14) N_{dcmac} times before being fed into a CNN block whereby data consistency is then applied afterwards. The application of equation (3.14) generates a data-consistent motion-augmented cine (DCMAC). In our study in the next section (section 3.5), we feed the CNN block with multiple inputs each of which have been generated with a range of $N_{\text{dcmac}} = \{0...30\}$. We also feed the CNN block, with an 'initial DCMAC' which is generated by applying a form of equation (3.14) to the acquired k-space data i.e. $x^0 = y$ (rather than the reconstruction from the previous cascade). We apply the equation successively several times, specifically, $N_{\text{initdcmac}} = 60$ times. The generation of these inputs to the CNN block is referred to as the DCMAC step. This is elaborated further

[2]or in other words, apply the DCMAC step N_{dcmac} times before every CNN denoiser bock

in the next section.

3.5 Exploiting Motion for Extremely Accelerated Cine MR Image Reconstruction

The aim of this section is to train a neural network to reconstruct $\mathbf{m}$ given an undersampled, zero-filled reconstruction $\hat{\mathbf{y}} = \mathcal{F}^{\mathrm{T}}\mathbf{y}$ by exploiting motion present in the cine MRI. By incorporating knowledge of the temporal dynamics into the reconstruction algorithm, data across all temporal frames in cine MRI sequence can be used to dealias any one particular frame.

We propose a novel deep learning approach for extremely-accelerated dynamic MR image reconstruction by exploiting motion present in MRI cines. The proposed method consists of three components: a motion estimation network; a *data-consistent motion-augmented cine* (DC-MAC) formed by intelligently propagating k-space information along the temporal axis; and a 3D CNN for MR image reconstruction. The use of the DC-MAC enables the incorporation of the full temporal k-space knowledge into the reconstruction algorithm, where data across the whole sequence can be utilised for dealiasing each frame. The network is trained end-to-end by minimising a composite loss function which consists of a motion estimation loss and an image reconstruction loss.

3.5.1 Methods

In brief, DC-CNN consists of N dealiasing units or 'cascades'. Each cascade takes a complex-valued estimate of the reconstruction as input (with additional 'data sharing' channels for neighbouring frames). It subsequently produces another, higher-quality cine as an output. This output cine is then subject to data consistency [93]. With our ME-CNN, we additionally provide each cascade with a novel *data-consistent motion-augmented cine*, also called x-DC-MAC, which exploits the full temporal information present in the original k-space data (with no data sharing required). As an additional set of channels, we also provide a method to motion-

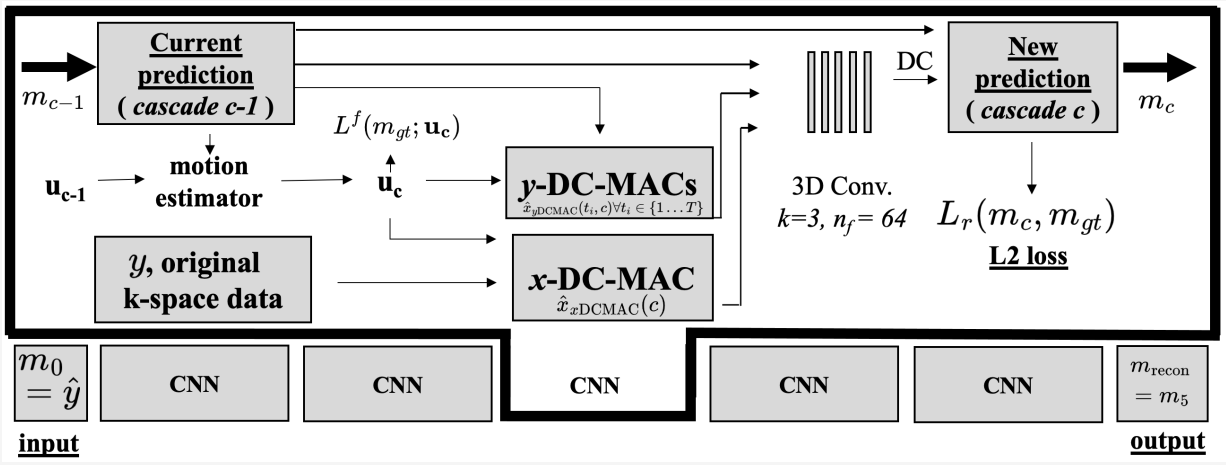

Figure 3.2: The proposed ME-CNN architecture with one example cascade illustrated. The motion estimator is based on the U-net architecture [68] as shown in Figure 3.3. The motion field, $\mathbf{u}_c$, is used to produce the DC-MACs which are then concatenating with the prediction from the previous cascade i.e. $\hat{m}_{c-1}$.

augmented the individual frame predictions from the previous cascade so that they can also be used for dealiasing. This is called y-DC-MAC. These DC-MACs are produced by learning a motion field (matrix M from section 3.3), for each cascade, c which is denoted as $\mathbf{u}_c$. The resulting process is illustrated in Figure 3.2 and forms the ME-CNN architecture.

In terms of training, each cascade, c, outputs two values - an optical flow representation and a predicted MRI reconstruction. These are used in the total loss function described in section 5 which consists of an optical flow loss and a reconstruction loss. In general, the predicted reconstruction, y_c, should improve in quality as you look at the output of deeper cascades. The prediction from the final cascade is used to produce the reconstruction loss. The optical flow, $\mathbf{u}_c^t$, of the MRI cine for each frame $t \in \{1...T\}$ is used to produce the optical flow loss. In our study $c \in \{1...N\}$ where $N = 5$ and $T = 30$ as thirty different cardiac phases were used in the construction of the dataset.

3.5.1.1 Motion Estimation

Within each cascade, a prediction of the motion field is made using an optical flow approach performed on the output of the previous cascade. In the case of the first cascade of the network,

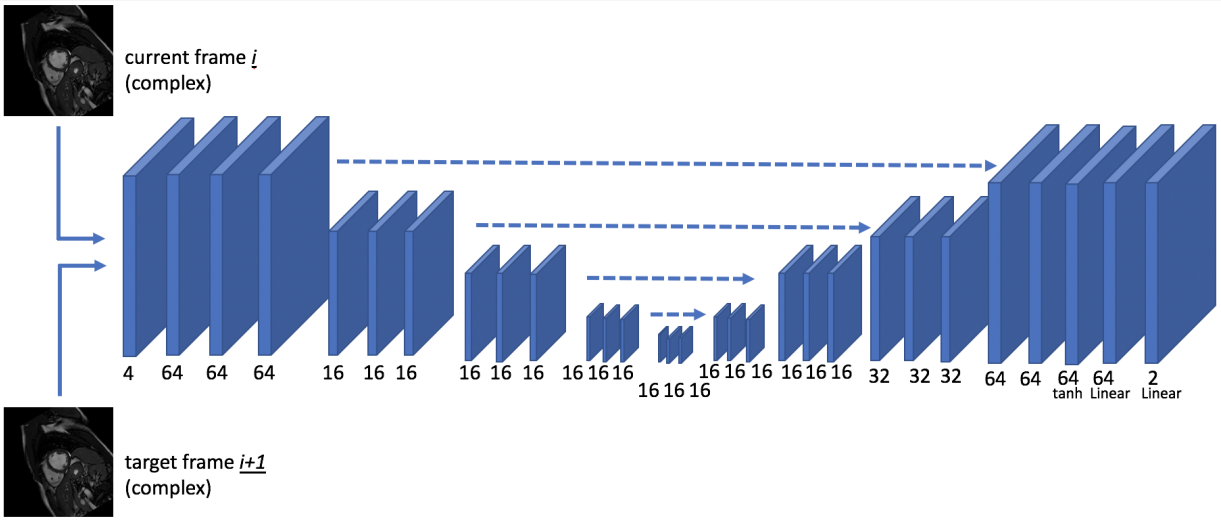

Figure 3.3: The motion estimator network used within each cascade of ME-CNN is illustrated. It is based on the U-net architecture with the addition of a convolutional layer at the start and end of the network. The input to this network is a pair of complex-valued images we wish to calculate the optical flow between. The feature map sizes for each convolutional scale in this network from start to end are $n_f = 64, 16, 16, 16, 16, 16, 16, 32, 64$ with tanh and linear activation functions for the final three layers

the original zero-filled reconstruction is used. [74, 114] showed that by training on the MSE loss between a motion-warped frame and its associated ground-truth, it is possible to learn the optical flow between frames in an unsupervised fashion. The important part of the loss for the motion field produced by a single cascade, c, is given by:

$$L_w^f(m_{\text{gt}}; \mathbf{u}_c) = \sum_{\mathbf{r},t} ||W(m_{\text{gt}}^t, \mathbf{u}_c^t) - m_{\text{gt}}^{t+1}||^2, \tag{3.16}$$

where m_{gt}^t is frame t of the ground truth, $\mathbf{u}_c^t$ is the motion field prediction from cascade c at time frame t and $W(m_{gt}^t, \mathbf{u}_c^t)$ warps frame m_{gt}^t using $\mathbf{u}_c^t$ and bilinear interpolation. There are two additional terms L_s^f and L_t^f which regularise the motion field, $\mathbf{u}_c$, with respect to its first order spatial and temporal gradients respectively. The total loss for the motion field output for a given cascade output in the network becomes equation (3.17) with hyperparameters α and β.

$$L^f(m_{gt}; \mathbf{u}_c) = L_w^f(m_{gt}; \mathbf{u}_c) + \alpha L_s^f(\mathbf{u}_c) + \beta L_t^f(\mathbf{u}_c). \tag{3.17}$$

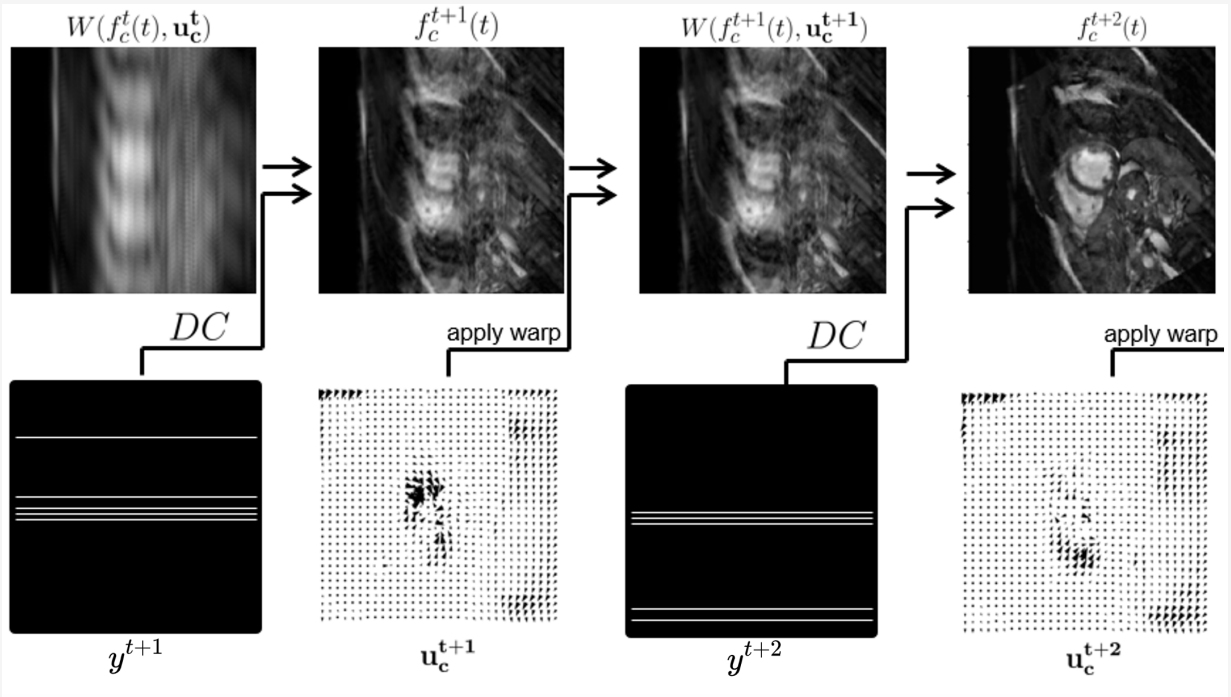

Figure 3.4: An example of how x-DC-MACs are produced for each cascade in ME-CNN. The first frame, $W(f_c^t(t), \mathbf{u_c^t})$, is the result of warping of frame $f_c^t(t)$ with $\mathbf{u_c^t}$. Since we are at the start of the process of generating the x-DC-MAC, $f_c^t(t)$ is simply equal to the zero-filled reconstruction from the original k-space information at time point t, i.e. $\hat{y}^t$. Data consistency is then applied to collect data from the next frame of the original k-space, y^{t+1}. This data is collected into the x-DC-MAC to generate the next frame in the sequence, $f_c^{t+1}(t)$. The process is then repeated using the motion field that warps from frame $t+1$ to frame $t+2$, $\mathbf{u_c^{t+1}}$.

3.5.1.2 Data-Consistent Motion-Augmented Cine (DC-MAC)

In order to better incorporate temporal information into the reconstruction of each frame, the full temporal-axis of the original k-space information is used to produce an intermediate cine reconstruction. The motion field, $\mathbf{u}_c$, from section 3.5.1.1 is used to propagate the acquired k-space information from one frame, t, to the next frame, $t+1$, whilst also acquiring the k-space information at frame $t+1$. This can be repeated iteratively for all subsequent frames until a k-space is achieved that has collected data from the entire temporal-axis of the original data. Given the acquisition data, $\mathbf{y}$ and a motion field, $\mathbf{u}$, the DC-MAC generation process is summarised by Figure 3.4. In particular, equation (3.18) shows the intermediate steps in the DC-MAC production process and equation (3.19) shows how the original k-space information, $\hat{x}$, is used in the generation of f_c^t, noting that x^t represents the zero-filled reconstruction for frame t.

$$f_c^{t+1}(t') = DC^{t+1} \circ W(f_c^t(t'), \mathbf{u}_c^t), \tag{3.18}$$

where DC^{t+1} is data consistency with the original k-space data, $\hat{x}$, at frame $t+1$ and t' is the frame number of the original k-space data to use as the first frame in the iterative warping process.

$$f_c^t(t) = \hat{y}^t \tag{3.19}$$

When using the initial condition set by equation (3.19), the desired DC-MAC, which is referred to as x-DC-MAC, using the motion field from cascade c is given by equation (3.20).

$$\hat{x}^t_{x\text{DCMAC}}(c) = f_c^{t+T}(t) \tag{3.20}$$

Using m_{c-1} to generate a set of DC-MACs for the reconstruction of m_c. In addition to the x-DC-MAC, we can use the output prediction of the previous cascade, m_{c-1}, with the motion estimation of the current cascade, $\mathbf{u_c}$, to make further warped projections of the output

cine. This is achieved by using a different initial condition from equation (3.19) within equation (3.18). Instead, $f_c^t(t) = m_{c-1}^t$ is used. The result is the generation of T additional predictive cines (since there are T frames to create different initial conditions from). This is summarised by $\hat{x}_{y\text{DCMAC}}^t(t_i, c) = f_c^t(t_i)$, where t_i is the frame number of the output from the previous cascade to use as the first frame in the iterative warping process.

3.5.2 Experiments

Architecture and comparison to DC-CNN The ME-CNN architecture is depicted in Figure 3.2. DC-CNN is trained with a data-sharing width $n_d = 5$ and feature map size of $n_f = 96$, giving a total of 3.9M parameters in the full network. Our proposed model uses $n_f = 64$ and introduces an additional motion estimation branch bringing the total network to 3.8M parameters. Like DC-CNN, ME-CNN uses residual connections between cascades.

The total loss function combining all cascades $c = \{1...N\}$ becomes:

$$L(x, \{m_1...m_N\}, \{\mathbf{u_1}...\mathbf{u_N}\}, m_{gt}) = \sum_{c}^{N} w_c(L_{\text{r}}(m_c, m_{gt}) + \gamma L^f(m_{gt}; \mathbf{u_c})), \tag{3.21}$$

where $w_i = 2^{-(N-i)}$ is the cascade-weight parameter. It is important to note that the motion estimation network produces predictions based on undersampled and intermediate network reconstructions m_c but is trained on warping ground-truth frames, m_{gt}^t.

DC-MAC vs Data Sharing (DS) The two mechanisms for sharing temporal information in DC-CNN are the convolutional layers and the data sharing mechanism. The proposed ME-CNN architecture replaces the data sharing mechanism with x-DC-MAC. The cine generated from data sharing with a depth of 5 frames [93] can be compared against x-DC-MAC output. Side-by-side in the video of Supplementary Material 1 in Appendix A.1 and here in Figure 3.5, it is clear that in spite of no deep learning dealiasing taking place, the x-DC-MAC is already of a much better quality compared to the DS cine. This further explains why our proposed ME-CNN would outperform that of the DC-CNN, particularly in high acceleration settings.

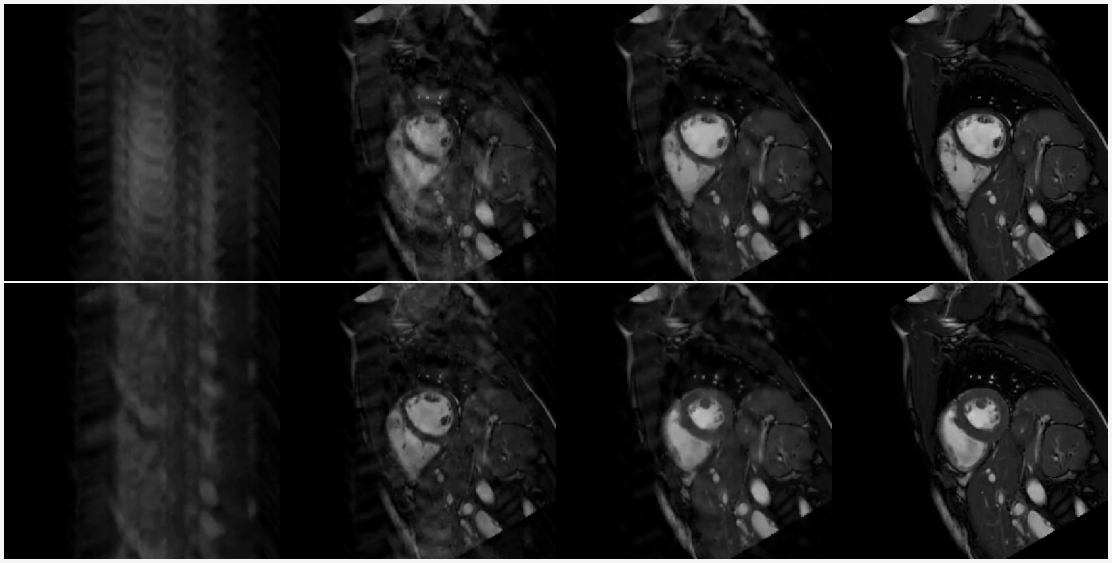

Figure 3.5: A comparison of Data Sharing (from DC-CNN) and DC-MAC. The top and bottom rows show the ED and ES frames respectively. From left to right: Zero-filled reconstruction at x16 acceleration, Data Sharing with a depth of 5, x-DC-MAC with 60 iterations, Ground Truth. It is clear that the temporal exploitation via the use of motion estimation is able to better preserve dynamic content of the cine.

Dataset and Data Augmentation In this study, the same dataset from the study in Schlemper *et al.* (2018) is used [93]. This consists of ten 256×256 short-axis cardiac MRI cines acquired with an SSFP sequence and $T = 30$ frames. The field of view of the scans was 320 × 320 mm and the thickness of the slice was 10 mm. In total, 32 coils were used in collecting the data. Normalising each coil against a body coil image, these coils were combined to generate a single coil complex-valued image of size 192 × 190 that was subsequently padded to 256 × 256. The associated single-coil k-space data was also used in order to supplement the data consistency layers present in our network and baseline. Seven scans are used for training, one for validation and two for testing. In order to help prevent overfitting and generalise the dealiasing process, the dataset was split into patches with a width of 32 pixels and retaining the original height of 256 pixels (which ensures that data-consistency can be applied). It was also augmented with on-the-fly random translations (±50 pixels), random rotations ($\pm 45^o$) and randomly generated Gaussian-centered variable density undersampling masks. For testing, we generated 1000 undersampling masks per test example resulting in a large augmented test set of 2000 cines.

Hyperparameters and configuration As advised in the study by Qin *et al.* (2018), we use

$\alpha = 1 \times 10^{-3}$, $\beta = 1 \times 10^{-4}$ [114] and $\gamma = 50.0$. We sampled 2 central lines in the k-space.

Training Each model was trained with a learning rate of 1×10^{-5} for 4×10^4 gradient steps, and then 1×10^{-6} for 2×10^4 gradient steps. This took 5 days on an NVIDIA Tesla P40 GPU but there are further advancements that will increase training speed in the near future. He initialisation was used with an Adam optimiser and parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$. The model was developed using the TensorFlow v1.8 Python API.

3.5.3 Results

Our model was evaluated using three-fold cross validation. On an initialised model, each cine took an average of 55 seconds to reconstruct on an NVIDIA Tesla P40 GPU with 24GB of memory. Table 3.1 shows the PSNR and SSIM statistics when computed across all three folds of the dataset. Across all three folds, ME-CNN performed better than DC-CNN with respect to SSIM for 100% of the test cases, and 89% of the time for PSNR. We also investigated the quality of the x-DC-MAC in the final cascade and found that there were around 27% of cases where the intermediate DC-MAC reconstruction produced better SSIM and PSNR than DC-CNN. As a result, given the DC-MAC, it is clear how the ME-CNN performs much better than DC-CNN. Figures 3.6-3.8 shows examples where ME-CNN have produced a perceptually better quality reconstruction. There are cases where the PSNR for DC-CNN are greater than that of ME-CNN. However, upon inspection, the ME-CNN still visually outperforms DC-CNN as is made clear by possessing a greater SSIM index (see Figure 3.9).

3.5.3.1 Using the motion field for reconstruction

The x-DC-MAC produced in this study was also examined. Using the model trained to reconstruct x51.2 undersampled cines, we evaluate our test set on much less aggressive undersampling rates of x9. Whilst the reconstruction outputs from DC-CNN were poor, ME-CNN produced more robust reconstructions. Furthermore, the x-DC-MAC produced from the final cascade of ME-CNN produced a quality greater than that of both DC-CNN and ME-CNN. This is

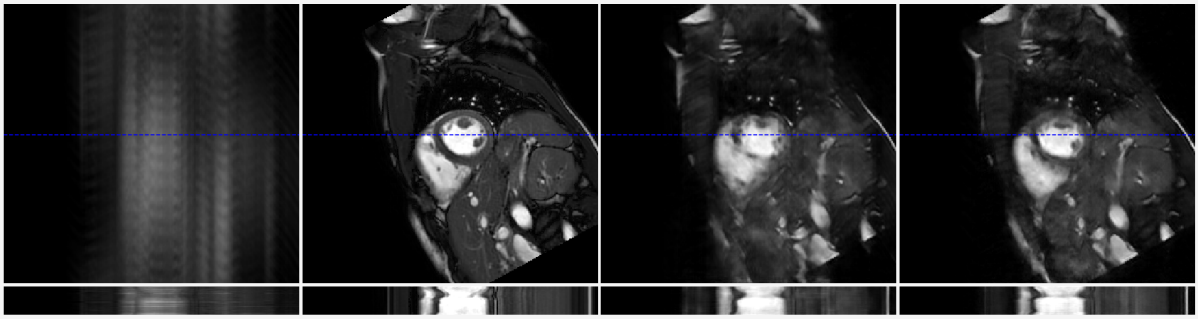

Figure 3.6: (a) x51.2 undersampled frame (b) Ground truth mid-motion frame (c) Baseline model. PSNR: 25.9, SSIM: 0.74 (d) Proposed model. PSNR: 27.7, SSIM: 0.80. The images on the bottom row shown the temporal variation (vertical axis) of the slice given by the blue dotted line. The cines for the full ground-truth, DC-CNN, ME-CNN and DC-MAC (from the final cascade of ME-CNN) are found in Supplementary Figures 2a, 2b, 2c and 2d respectively (see appendix A).

shown in Figure 3.10. The motion field was used to warp each frame to the next frame in the ground-truth cine sequence, thus generating a new cine upon which motion field quality can be partially determined. For the x51.2 experiment, the cine produced an average PSNR of 39.5 ± 1.9 which is comparable to that of the x9 experiment of 39.6 ± 1.8 which indicates that whilst the dealiasing part of the network has not generalised well to other undersampling rates, the motion estimation network has. The generalisability of the motion estimation arguably helps provide more robustness to the neural network to unseen examples. Indeed, the perceptual quality of the reconstructions from ME-CNN outperform that of DC-CNN as shown in table 3.2. Further experiments are required to see if there exists domain shifts where ME-CNN doesn't outperform DC-CNN.

Table 3.1: A comparison of the reconstructions produced by 3 different models, DC-CNN, DC-MAC and ME-CNN. The DC-MAC is from the final cascade of ME-CNN, $y^t_{x\text{MAC}}(N)$. The difference in performance of each method on the same test sample is also recorded with the mean difference given by the entries starting with a 'Δ'.

Model	PSNR	SSIM
DC-CNN	24.4 ± 2.4	0.670 ± 0.081
DC-MAC	23.0 ± 1.8	0.628 ± 0.041
ME-CNN	$\mathbf{27.3 \pm 2.5}$	$\mathbf{0.776 \pm 0.054}$
Δ(ME-CNN − DC-CNN)	$\mathbf{2.82 \pm 2.34}$	$\mathbf{0.106 \pm 0.079}$
Δ(DC-MAC − DC-CNN)	-1.42 ± 1.77	-0.042 ± 0.062

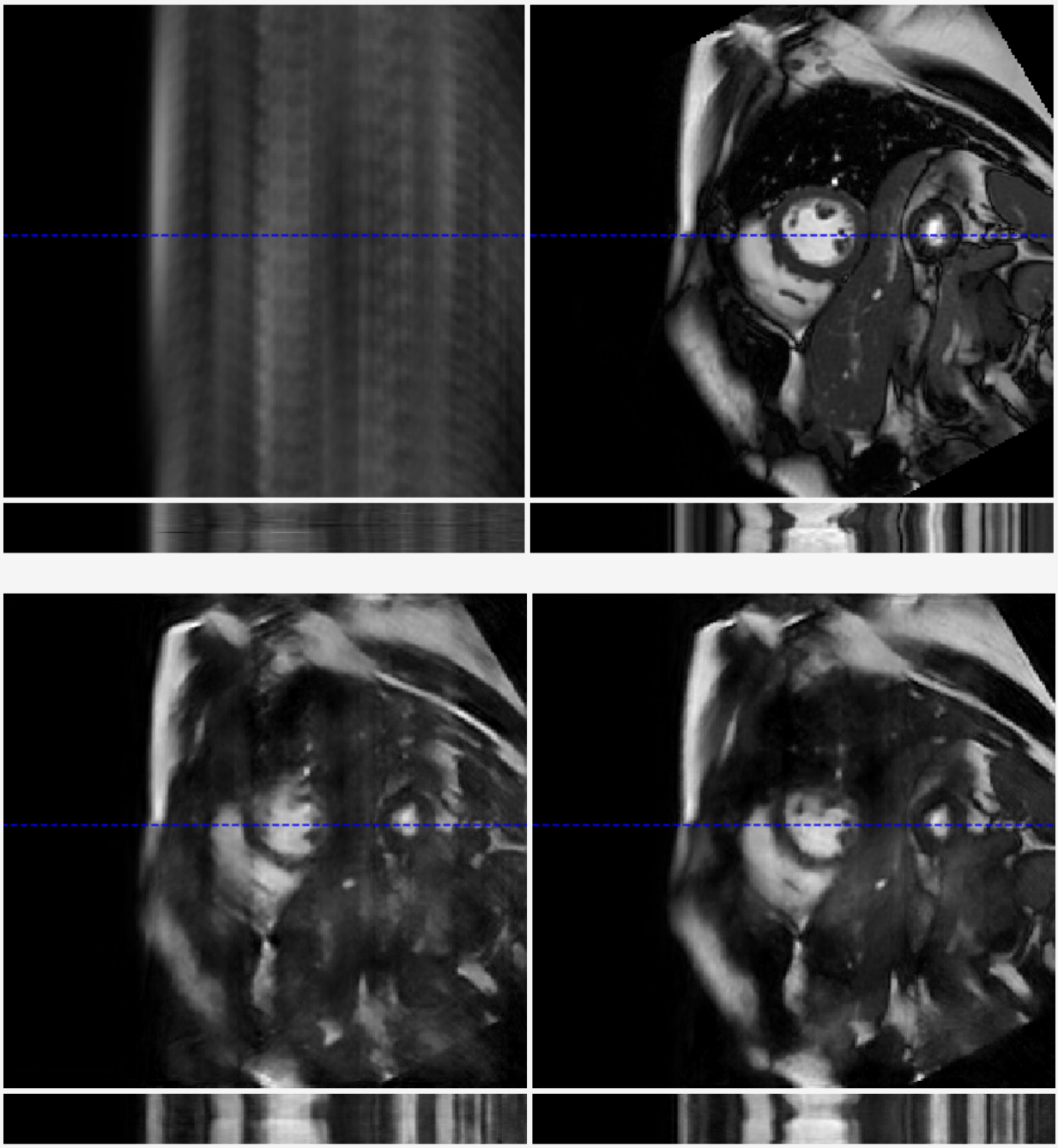

Figure 3.7: Another example of better quality being produced by ME-CNN compared to DC-CNN. Clockwise from the top-left image: (a) Undersampled input cine (b) Ground truth of a mid-motion frame, with the blue dotted line indicating the slice used to show the temporal variation in the bottom row (c) DC-CNN output. PSNR: 25.21 SSIM: 0.72 (d) ME-CNN output. PSNR: 26.96 SSIM: 0.79. The ground truth, DC-CNN reconstruction, ME-CNN and DC-MAC reconstruction cines are found in the MP4 files labelled Supplementary Figures 3a, 3b, 3c and 3d respectively (see appendix A).

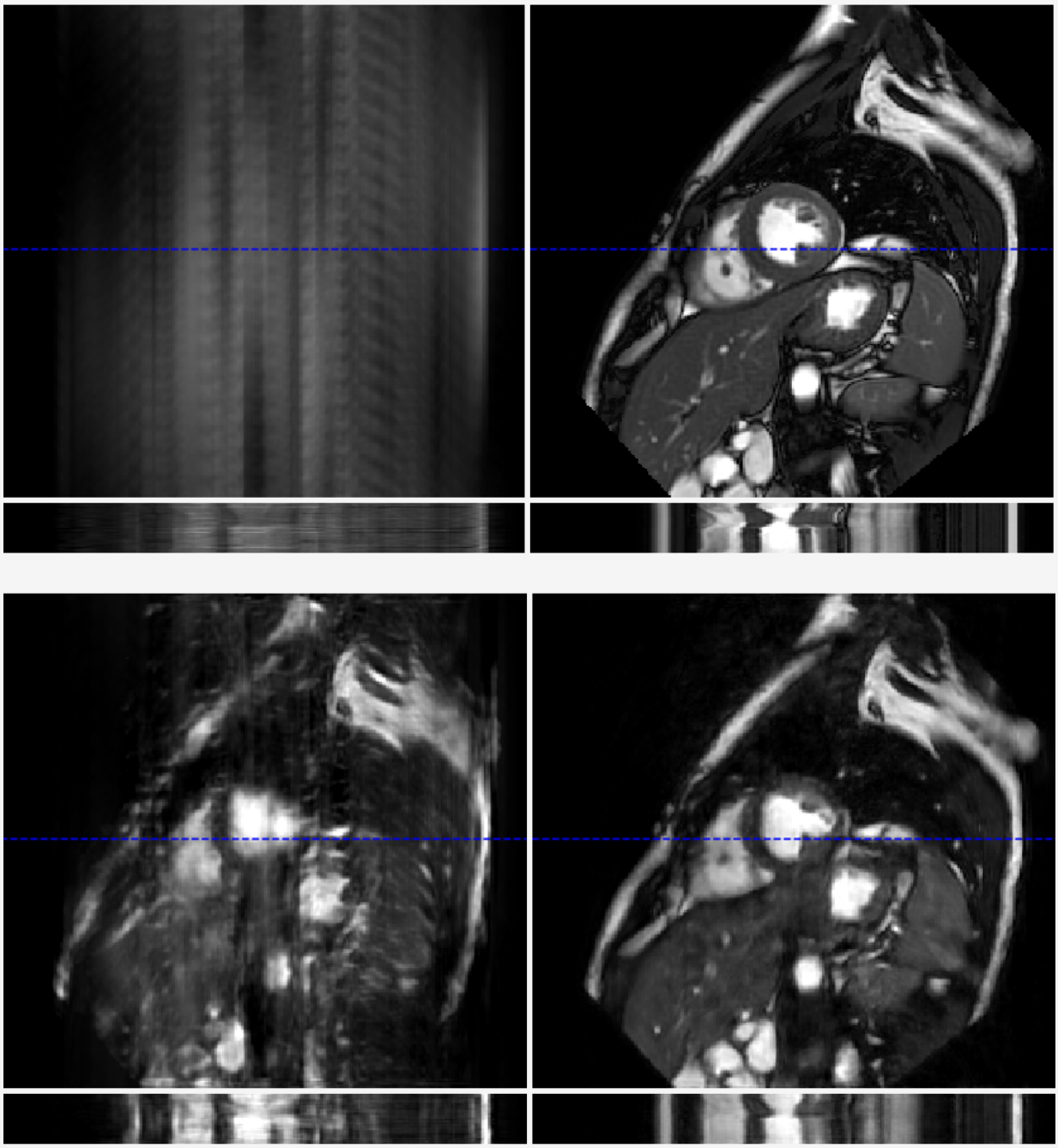

Figure 3.8: Another example of better quality being produced by ME-CNN compared to DC-CNN. Clockwise from the top-left image: (a) Undersampled input cine (b) Ground truth of a mid-motion frame, with the blue dotted line indicating the slice used to show the temporal variation in the bottom row (c) DC-CNN output. PSNR: 22.6 SSIM: 0.57 (d) ME-CNN output. PSNR: 27.1 SSIM: 0.76. The ground truth, DC-CNN reconstruction, ME-CNN and DC-MAC reconstruction cines are found in the MP4 files labelled Supplementary Figures 4a, 4b, 4c and 4d respectively (see appendix A).

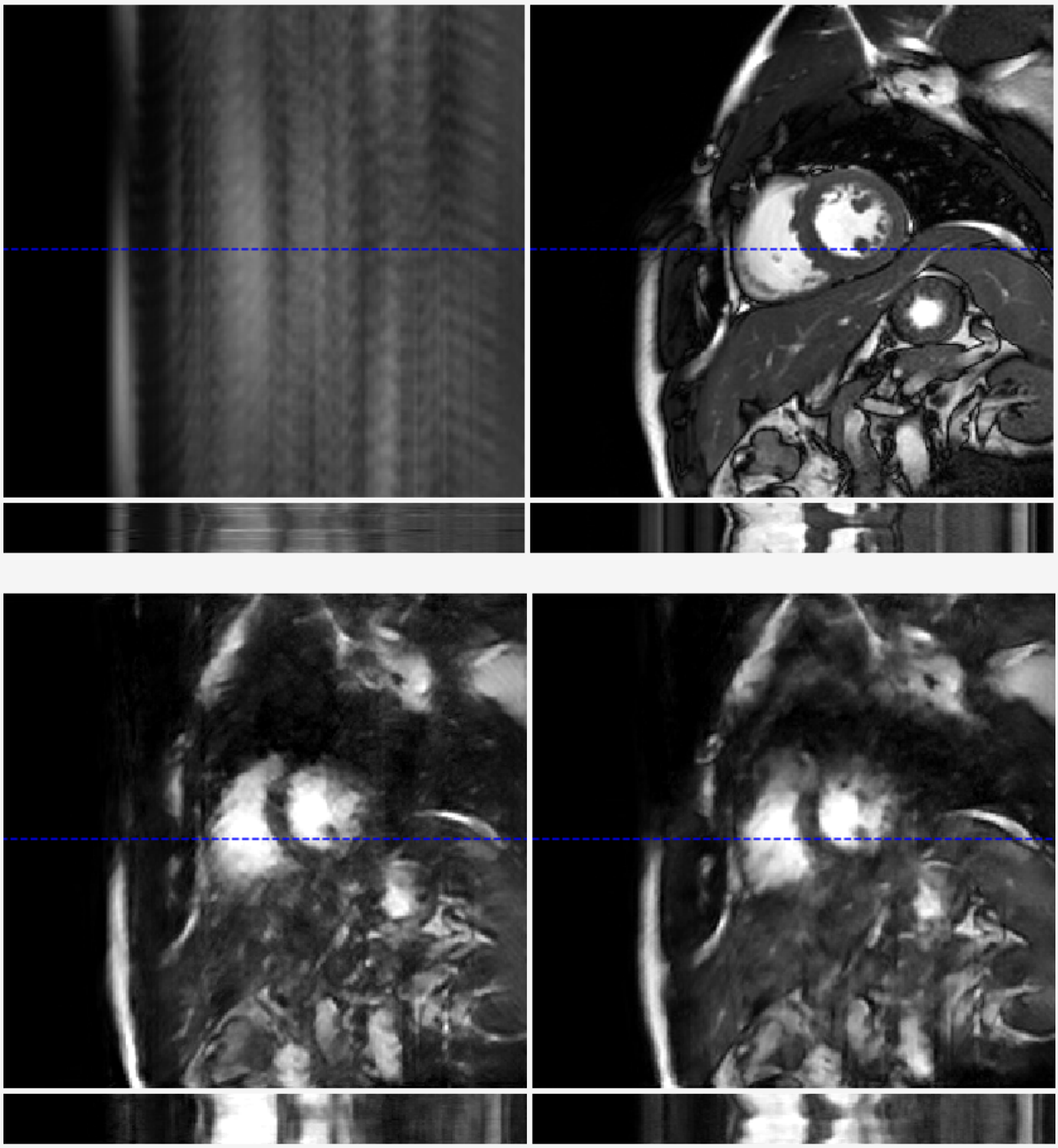

Figure 3.9: An example where DC-CNN outperforms ME-CNN with respect to PSNR but still produces poor spatiotemporal quality in comparison to ME-CNN as indicated by the lower SSIM index. Clockwise from the top-left image: (a) Undersampled input cine (b) Ground truth of a mid-motion frame, with the blue dotted line indicating the slice used to show the temporal variation in the bottom row (c) DC-CNN output. PSNR: 23.1 SSIM: 0.65 (d) ME-CNN output. PSNR: 23.0 SSIM: 0.67. The ground truth, DC-CNN reconstruction, ME-CNN and DC-MAC reconstruction cines are found in the MP4 files labelled Supplementary Figures 5a, 5b, 5c and 5d respectively (see appendix A).

Table 3.2: Video quality metrics when testing the generalisability of the trained x51.2-acceleration reconstruction models on x9-accelerated test data.

Model	PSNR	SSIM
DC-CNN	21.9 ± 5.3	0.570 ± 0.190
DC-MAC	$\mathbf{34.3 \pm 1.9}$	$\mathbf{0.930 \pm 0.023}$
ME-CNN	31.5 ± 1.5	0.874 ± 0.019
Δ(ME-CNN − DC-CNN)	9.57 ± 5.44	0.305 ± 0.182
Δ(DC-MAC − DC-CNN)	$\mathbf{12.44 \pm 5.45}$	$\mathbf{0.361 \pm 0.188}$

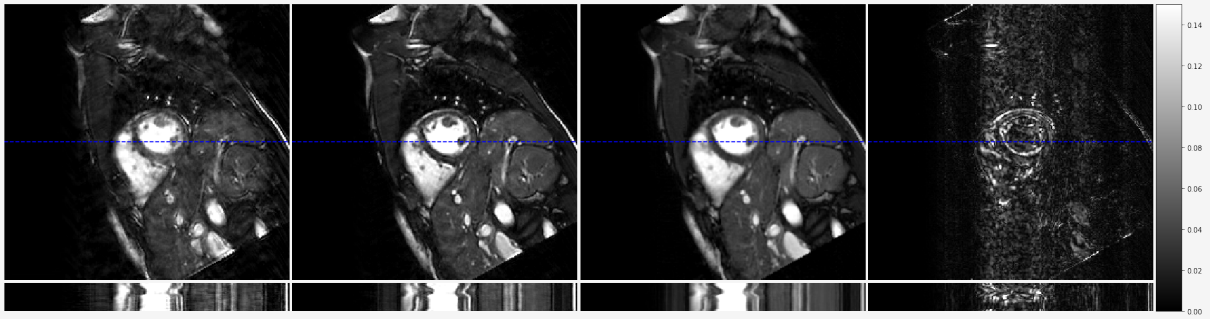

Figure 3.10: (a) Baseline. PSNR: 28.29, SSIM: 0.75 (b) ME-CNN. PSNR: 32.2, SSIM: 0.89 (c) x-DC-MAC. PSNR: 35.2, SSIM: 0.94 (d) Absolute difference between x-DC-MAC and ground truth with colorbar (cine dynamic range is 1.0). The ground truth image is found in Figure 3.6. The full ground-truth, x-DC-MAC for the final cascade and the absolute error cines are found in Supplementary Figures 6a, 6b and 6c respectively (see appendix A).

3.5.4 Conclusion

A notable remark on future directions of the ME-CNN involves the generation of the DC-MAC reconstructions. The generation of the DC-MACs used T number of motion field warps. In theory, you could perform another T iterations to better incorporate temporal information. Finally, it is worth noting that similar concepts involving motion estimation have been used for super-resolution applications [76]. With modifications to the x-DC-MAC and y-DC-MAC production, super-resolution (SR) MRI with ME-SR-CNN remains a possibility.

In conclusion, we observe that for aggressive undersampling rates where the DC-CNN approach is not able to effectively exploit the full temporal domain, the proposed ME-CNN is able to

outperform the state-of-the-art approach. The proposed end-to-end trainable model was able to generate motion field estimations which were used to produce the DC-MACs. The increased performance of the network is largely due to the DC-MACs which allow the entire temporal axis to be exploited. We have demonstrated that ME-CNN is a more robust approach to dealiasing unseen examples of different acceleration rates. For future work, we will explore the use of ME-CNN for 3D reconstructions. We will also modify the ME-CNN to explore potential real-time imaging applications.

3.6 Robust Dynamic MRI Reconstruction for Active Acquisition pipelines

3.6.1 Introduction

Recent works have explored the use of active acquisition of MRI k-space for image reconstruction with the aim of reducing scan time without compromising or being technically-limited in obtainable image/cine quality [147]. With the advent of the these new acquisition-reconstruction pipelines, we show that motion-exploiting reconstruction networks can produce reconstructions that are robust to the pipeline's choice of undersampling mask.

In a study investigating active acquisition [147], a ResNet-based architecture with similarities to the DC-CNN [93] is used for reconstruction. Their 'cResNet' reconstruction network produces an 2D static image which is evaluated by a separate network to decide which line to next acquire to reduce the overall uncertainty in the reconstruction. In this study, we argue that for dynamic MRI, a better reconstruction network would be the recent ME-CNN (motion-exploiting convolutional neural network) [141] which exploits the dynamic nature of the data to harness the entire k-space budget available. Additionally, we study a key component of the ME-CNN, the DC-MAC (data-consistent motion-augmented-cine).

3.6.2 Method

The ME-CNN resembles the DC-CNN except that each CNN/cascade contains a motion estimation block. The motion estimation block is used to generate a DC-MAC that is appended onto the CNN to aid in dealiasing [141].

The DC-MAC can be formulated as a temporally-iterative noiseless data-consistency term:

$$r^{t+1} = \mathcal{F}^{-1}\Big[(D^{t+1} \circ y^{t+1}) + ((1 - D^{t+1}) \circ \mathcal{F}(M^t r^t))\Big] \tag{3.22}$$

In this case, the matrix M^t is the bilinear interpolation operation that warps its subject, r^t, with an optical flow $\mathbf{u^t}$ which represents the motion from frame t to $t+1$. Note, the iterative algorithm is dependent on the initial frame for reconstruction, r^0, which is set to a frame, t' from the zero-filled reconstruction, $\hat{y}^{t'}$. $\hat{x}_{\mathrm{x}}(t)$ is the value of r^{nT} when zero-fill reconstructed frame $\hat{y}^t$ is used as the initial frame r^0, where n is an integer. We set $n = 2$ in this study after a preliminary investigation with our fully-sampled dataset and the motion-field learned by a U-Net [114, 74, 68] shows that the DC-MAC is optimised in terms of the SSIM and PSNR. One possible reason why larger n causes PSNR and SSIM to drastically drop for high acceleration rates is because a k-space residue builds up in the regions of the k-space that are sampled less-frequently or not sampled at all.

The previously mentioned study uses the cResNet, a cascading FC-ResNet-based method to dealias static 2D images [106, 147]. The architecture draws inspiration from the DC-CNN. To reproduce the cResNet experiment, it would have taken 187M parameters which may not provide a fair comparison to ME-CNN with only 3.2M parameters. When training with a reduced parameter model of cResNet (achieved by using fewer filters), the model heavily overfitted to our dataset. Instead, the DC-CNN was used as its closest model that performed reasonably. The DC-CNN consists of a series of N-cascades of CNNs. Each CNN contains 5 convolutional layers, $n_f = 96$, $k = 3$ with ReLU non-linearity. Each CNN is followed by the application of data-consistency which ensures that the reconstruction prediction made by the CNN is consis-

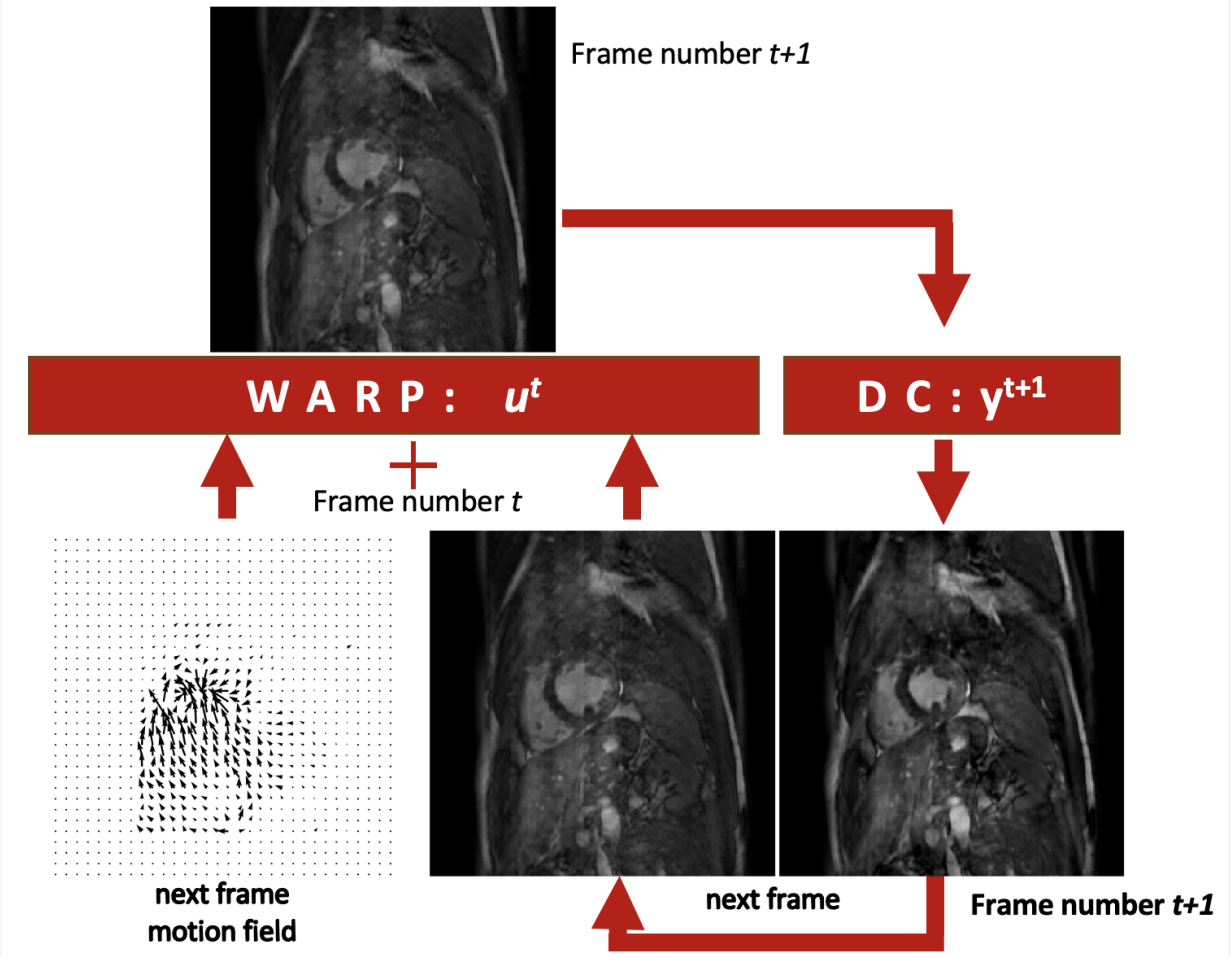

Figure 3.11: Diagram explaining the DC-MAC generation process. Given an optical flow, u^t, v^t, that propagates a frame from frame t to $t+1$, a zero-fill reconstructed frame can be warped by the flow using bilinear interpolation. The resulting frame can be then made to be data-consistent with the original k-space data at frame $t+1$, y^{t+1}. The resulting frame has now collected the part of the k-space budget for frame $t+1$. This new frame becomes the new initial frame in this process.

tent with the originally acquired k-space data [93, 141].

3.6.2.1 Training

We train the ME-CNN end-to-end with an L2 loss only on the output reconstruction of the final cascade in the network and with $N = 3$ cascades, different to the original design. Additionally, we train the optical flow output of each cascade, C, against a L2 warp loss, L^f with hyperparameter $\gamma = 15.0$. The final loss function becomes (3.23).

$$L(m, m_{gt}) = ||m - m_{gt}||^2 + \gamma \sum_{c}^{n_c} L^f(m_{gt}; \mathbf{u_c}) \tag{3.23}$$

3.6.3 Dataset

We use ten short-axis cardiac cines obtained using SSFP acquisition, 320×320mm field-of-view, 10mm slice thickness. 32 channels were obtained and combined using SENSE reconstruction [16] to generate a single CMR cine. During training, we augment the dataset with a random undersampling mask, random rotation and translation. The undersampling mask generated follows a Gaussian distribution, centered in the middle of k-space with anywhere between 3 and 85 lines acquired per frame (uniformly distributed).

3.6.4 Results

We evaluated the models on acceleration rates from ×3 to ×125 and found that ME-CNN on average performed better than DC-CNN in terms of both PSNR and SSIM as can be seen in Figure 3.13. In particular, the average difference in performance was higher for the ME-CNN compared to the DC-CNN. For high acceleration rates, the standard deviation of this difference was lower than the average at the corresponding acceleration rate. This implies a high statistical significance for claims that ME-CNN outperforms DC-CNN. An example reconstruction can be found in Figure 3.12.

For lower acceleration rates, this claim is weaker as the average difference is lower and the standard deviation of the difference increases. However, it should be noted that for low acceleration rates, the average difference in performance for the DC-MAC compared with the DC-CNN is greater than the average difference for the ME-CNN compared with the DC-CNN. The robustness of the DC-MAC at these low acceleration rates leads to a low standard deviation in the difference which means that high statistical significance can be found for the claim that DC-MAC outperforms DC-CNN at these rates.

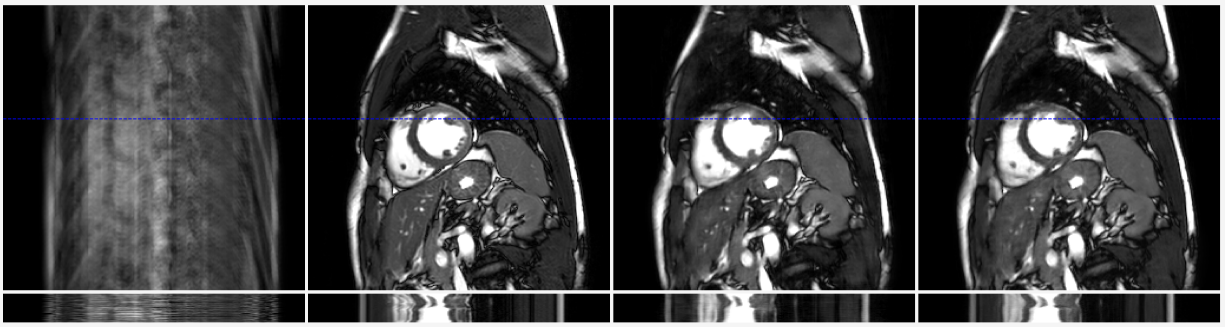

Figure 3.12: An example reconstruction with acceleration rate x16. Left-to-right: Zero-filled reconstruction, DC-CNN output, ME-CNN output, Ground Truth. PSNR, SSIM respectively for this example: DC-CNN 35.10, 0.93, ME-CNN 36.87, 0.95.

3.6.5 Conclusion

In this study, we present findings that demonstrate the robustness of ME-CNN and the associated DC-MAC to MRI acquisitions with a random number of acquired lines in comparison to DC-CNN which closely resembled the cResNet. The DC-MAC is able to act as an intermediate reconstruction that not only stabilises the output of the ME-CNN but for high undersampling rates, is able to extract knowledge from the entire k-space budget. The DC-CNN is limited by its temporal receptive field and learn to stabilise its reconstruction against the random nature of the undersampling mask. It should be noted that this work was conducted for 2D+t cardiac imaging. Future work includes extending our method to other anatomy and for 3D+t imaging.

3.7 ME-CRNN: Using CRNNs for unrolled optimisation of Motion-based MRI reconstruction

In the previous section, cine reconstruction was made possible by the optimisation of equation (3.3) by only intermittently applying the denoising step of equation (3.15). Whilst this produced competitive results in an extremely accelerated setting, it is suboptimal to skip too many of these denoising steps. In this section, we study the case in which this denoising is applied at

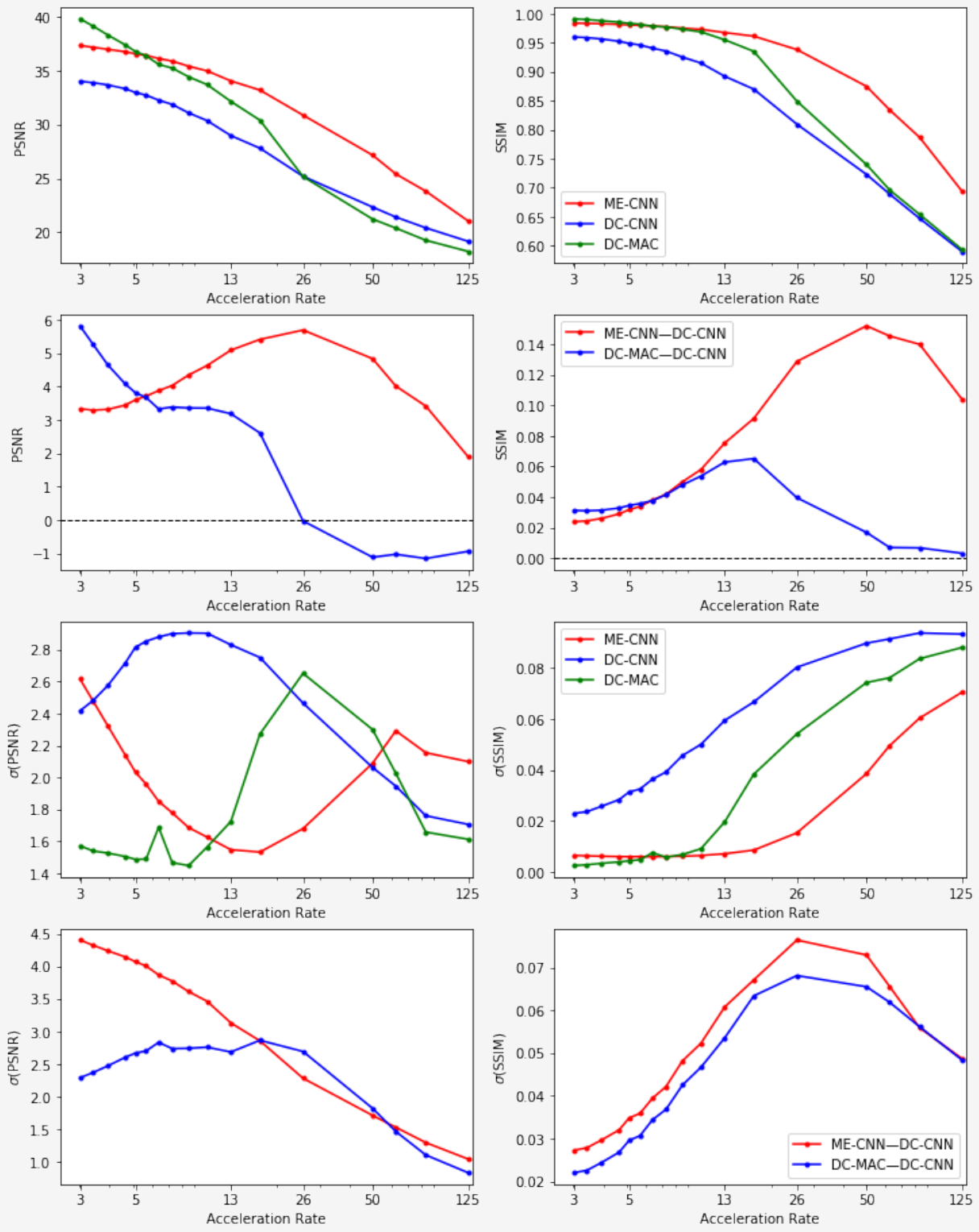

Figure 3.13: Results comparing the ME-CNN, DC-MAC and DC-CNN reconstruction with acceleration rates from ×3 to ×125. Left column: PSNR metric. Right column: SSIM metric. First row: the average performance for each acceleration rate. Second row: the average of the difference in the performance metric. Third row: the standard deviation of the metric. Fourth row: the standard deviation of the difference in the performance metric. Note the log scale on the x-axis.

every iteration.

Noting that the denoising term is parameterised by a deep neural network, traditionally N iterations of this approach would require $N(\theta + \phi)$ parameters where θ and ϕ are the number of parameters in the denoiser and motion estimator respectively. For more than 5 iterations, fitting this into typical GPU memory in a time-efficient manner is infeasible mainly due to the need to also train the motion estimators alongside the denoisers.

In order to simplify the problem, we use a single, trained motion estimator for all iterations. The input to the dealiasing network includes this evaluated motion estimate which is generated directly from the undersampled zero-filled reconstruction. This motion estimator network is trained separately using an optical flow loss and a smoothing term on the motion estimate as in equation (3.16).

One particular choice for the motion estimator include the state-of-the-art VoxelMorph which can be extended to the cascading case and for use with undersampled images [122, 148]. However, instead we use a 3D convolutional autoencoder-like network which may fail to capture higher resolution motion details. The autoencoder-like network contains 4 downsampling layers deep with feature maps sizes of 64, 128, 256 and 512 generating a latent space of 262k neural units as depicted in Figure 3.14. By using this crude motion estimator, we hope to demonstrate the robustness of motion-based image reconstruction even when motion estimates are suboptimal. For example, out-of-training-domain cases for particular cardiac views may generate optical flows with too much or too little displacement. It should be noted that there exists a rich literature to produce motion estimates of higher quality than this autoencoder-like network. This offers opportunities for future investigations of motion estimator choice for the purpose of optimising reconstructions, particularly when the motion estimate is not jointly trained with the reconstruction [74, 114, 42, 87, 94].

We also evaluate our method using some higher detail motion estimates which are produced from the ME-CNN network from the previous section [141] that performs 5 iterations of joint reconstruction-motion-estimation optimisation. We denote these experiments with '(HQ MF)' to indicate that a reconstruction-based motion estimate is used.

Further to this, to reduce the number of parameters in the model, we re-use the model weights by utilising an RNN framework [113]. Each RNN cell represents a proximal operation that acts in favour towards the objective in equation (3.8). The advantages of fewer weights in an RNN framework includes reduced overfitting and increased training/test-time speed. In addition to this, if future work were to include a motion update term at each iteration, then the RNN framework would be essential due to the typically large memory requirements of performing motion estimation.

We perform the experiment using 10 and 60 iterations for an acceleration rate of x16 with variable density Cartesian undersampling. A set of control (CNTL) experiments are also evaluated which do not use any motion term in its iterative reconstruction. A summary of the ME-CRNN is depicted in Figure 3.15.

3.7.1 Dataset: UK BioBank

The UK BioBank (UKBB) consists of data from 500,000 volunteers aged between 40 and 69 with the primary focus on studying adult diseases. The participants were enrolled between 2006 and 2010 with continued monitoring. The dataset includes genetic data in order to explore the potential link with certain diseases. Alongside this is a wealth of other data such as brain, heart and full body imaging, biochemical markets, questionnaires and blood samples [69]. At the time of writing, over 25,000 subjects have had cardiac MR (CMR) scans which includes multiple types of acquisitions [57]. In particular, multiple slices of long and short axis (views of the heart) cines are acquired which are extensively used in this thesis. The UKBB is a remarkable achievement that has helped bridge the gap between theoretical research (particularly computer scientists) and population impact [105]. Studies involving the UKBB are ongoing with updates to the acquisition protocols being recommended to increase the quality of the dataset [173, 105].

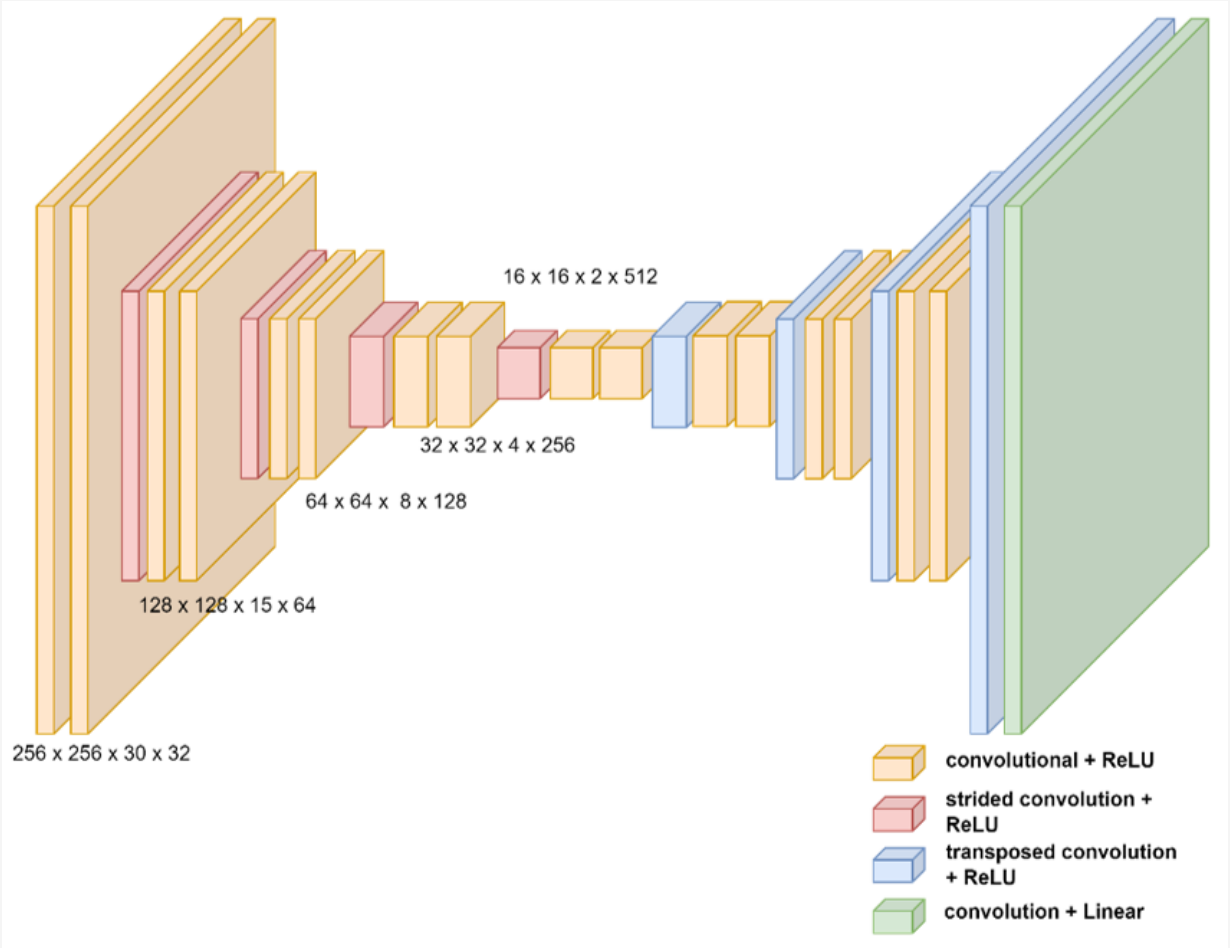

Figure 3.14: The autoencoder-like network used to obtain crude motion estimate for the ME-CRNN study. The dimensions of each layer are shown in the format of width × height × time × features/channels.

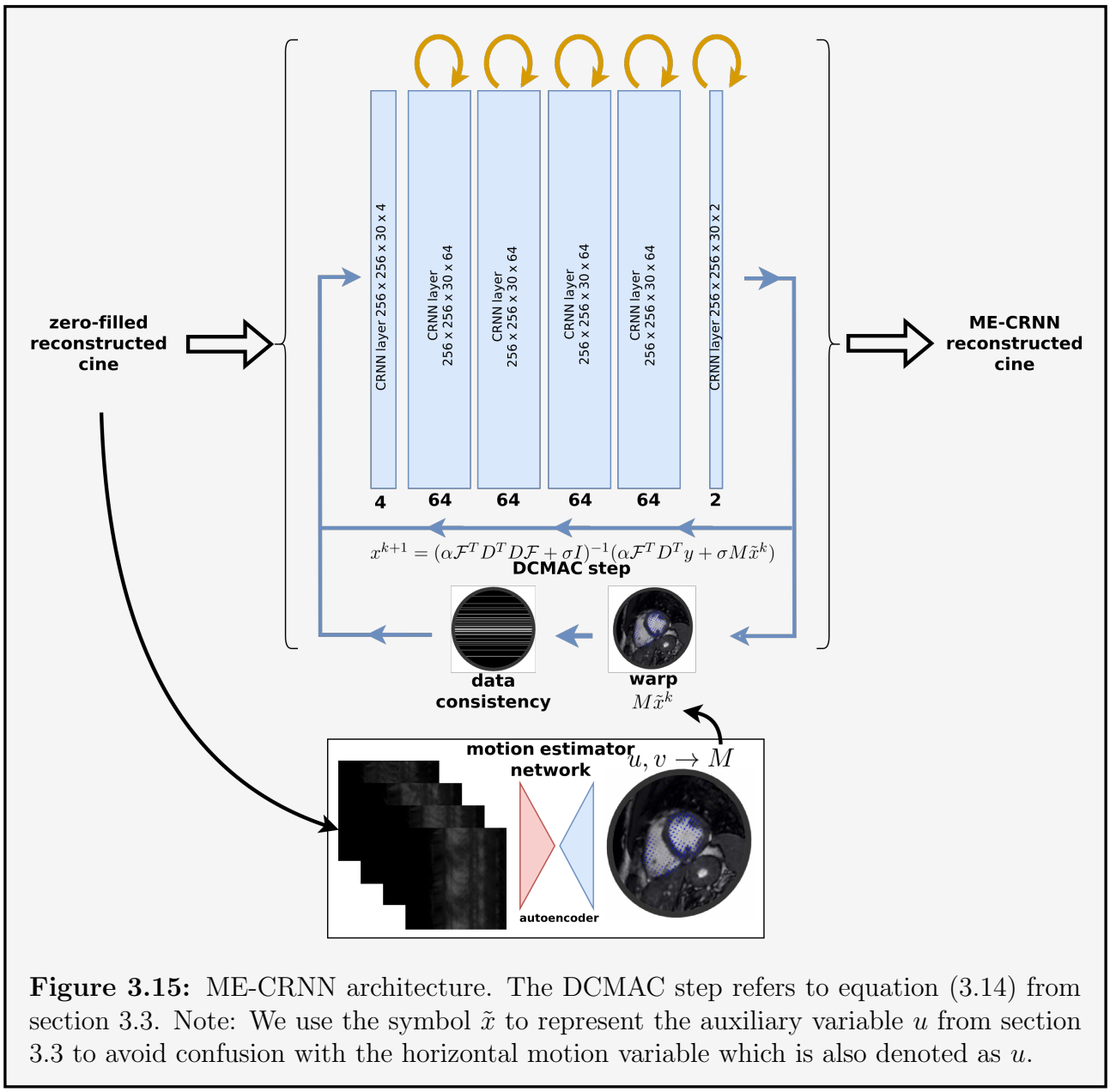

Figure 3.15: ME-CRNN architecture. The DCMAC step refers to equation (3.14) from section 3.3. Note: We use the symbol $\tilde{x}$ to represent the auxiliary variable u from section 3.3 to avoid confusion with the horizontal motion variable which is also denoted as u.

3.7.2 Results

We evaluated various models on cardiac data from [93] and the UK BioBank study with synthetic phases:

1. *CRNN* [CNTL] - This is a suitable baseline and matches the implementation from the study in [113] except without bidirectional temporal recurrent units and using 3D layers instead of 2D
2. *ME-CRNN* - This is our proposed model which can be appropriately compared with CRNN as the control.
3. *2D MECRNN* - This is a version of ME-CRNN but using 2D layers instead of 3D layers. There is no bidirectional temporal current unit. The memory gain allowed us to fit 60 iterations instead of just 10, however, the gradients were only applied to the last ten iterations during backpropagation.
4. *2D BiCRNN* [CNTL] - This is the model from [113].
5. *ME-BiCRNN* - This is a version of the ME-CRNN but using only 2D layers, not 3D, and using bidirectional recurrent temporal units from the 2D BiCRNN.
6. *kt-FOCUSS* - This is a classical reconstruction algorithm that exploits sparsity in the $x-f$ domain of the acquisition. A version that incorporate motion compensation is denoted with "ME/MC" that requires a fully sampled reference is also included in these results. See 4.1.3 for more information.

The results of these studies can be found in tables 3.3 and 3.4. Figures 3.16-3.21 are some results from the evaluation.

3.7.3 Discussion

The results show a clear edge of ME-CRNN and CRNN over the more traditional approach, kt-FOCUSS. The training dataset regularised the CNN denoisers providing a distinct advantage

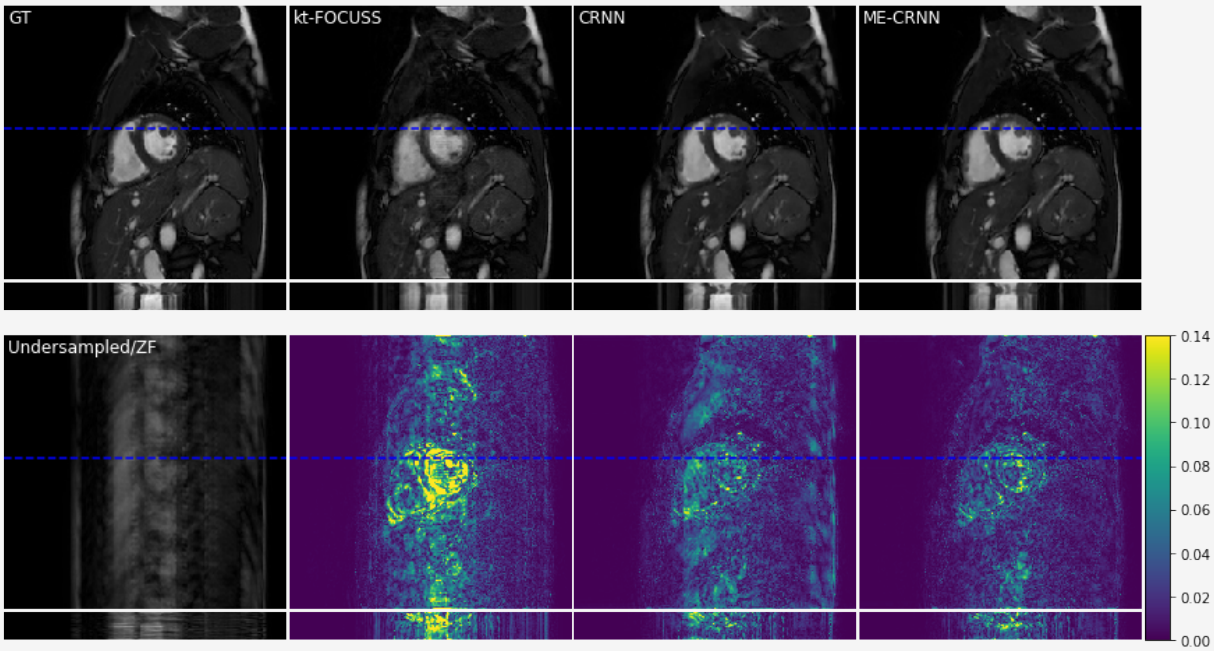

Figure 3.16: Comparison of the reconstruction of a x16 cardiac cine acquisition by various models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: kt-FOCUSS - (32.5, 0.914), CRNN - (34.8, 0.940), ME-CRNN - (36.8, 0.956).

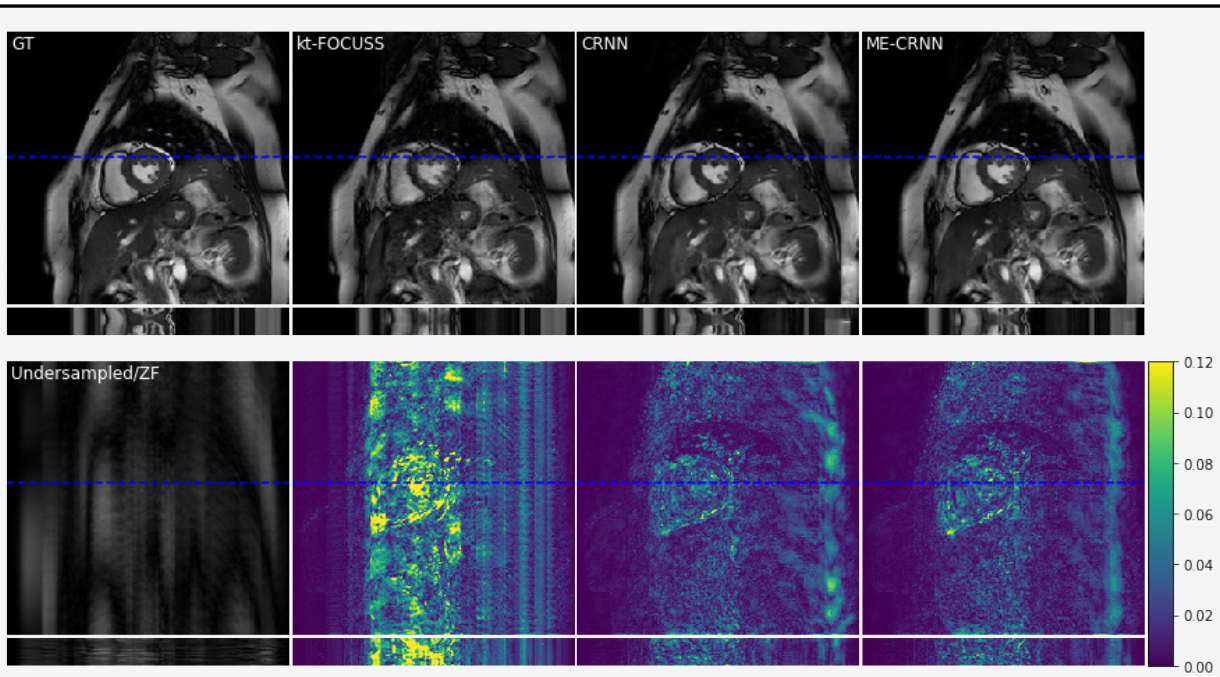

Figure 3.17: Comparison of the reconstruction of a x16 cardiac cine acquisition by various models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: kt-FOCUSS - (29.6, 0.891), CRNN - (28.3, 0.899), ME-CRNN - (32.6, 0.931).

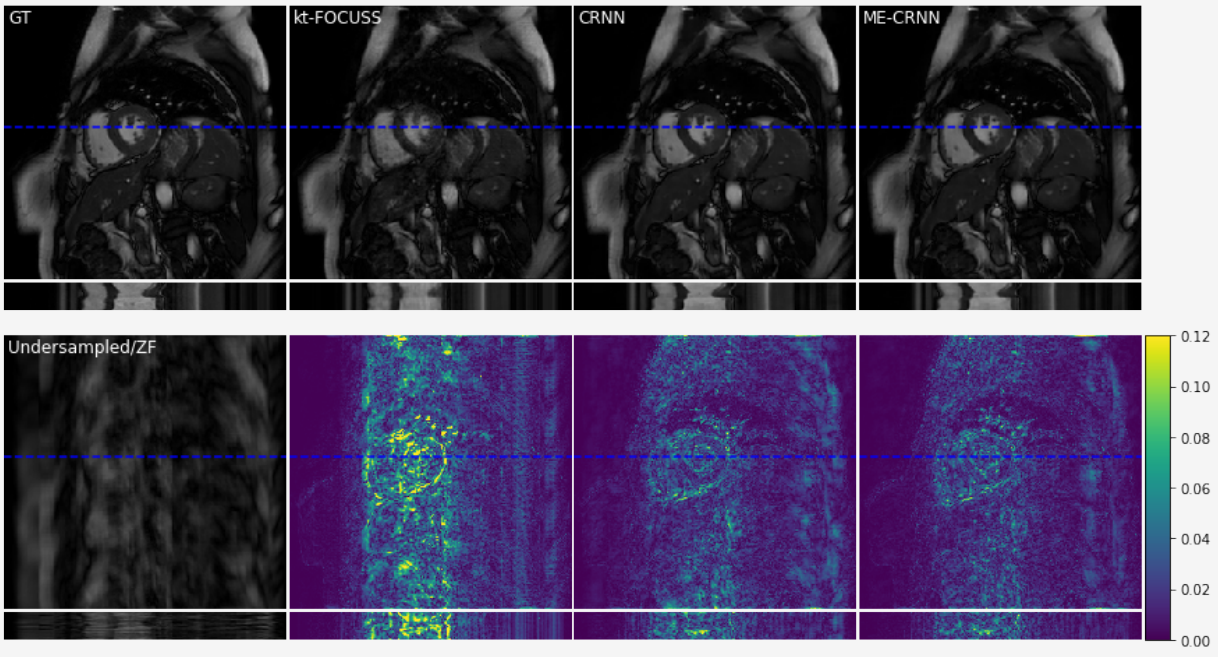

Figure 3.18: Comparison of the reconstruction of a x16 cardiac cine acquisition by various models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: kt-FOCUSS - (32.3, 0.898), CRNN - (35.9, 0.940), ME-CRNN - (37.2, 0.954).

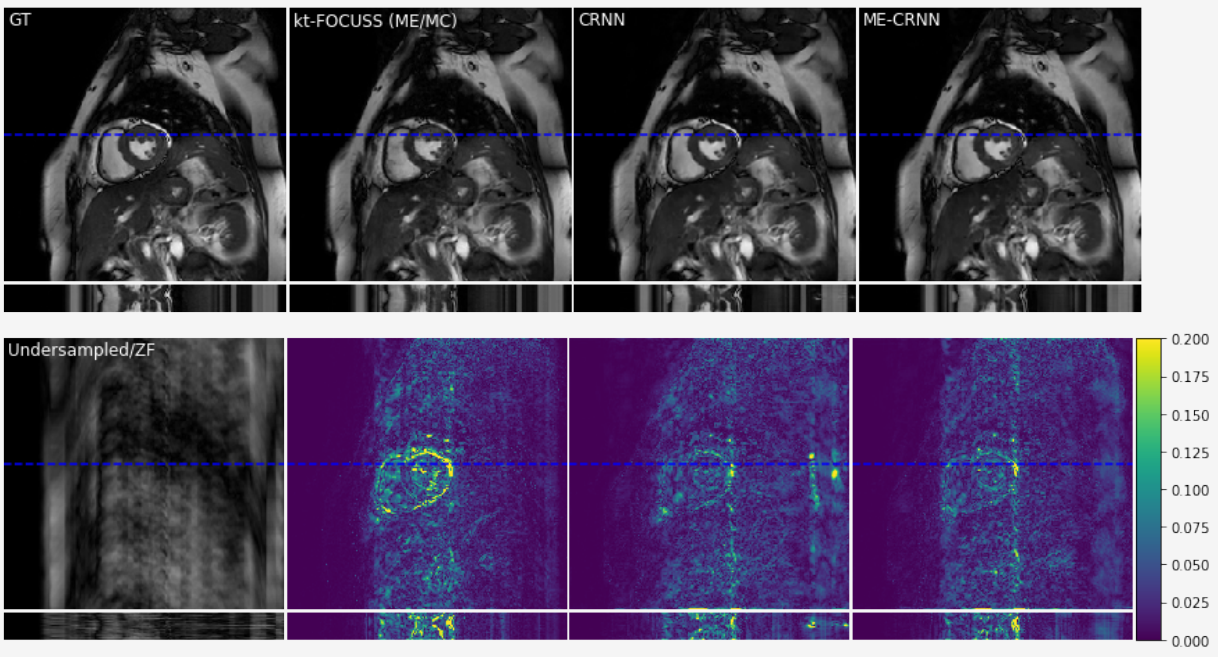

Figure 3.19: Comparison of the reconstruction of a x16 cardiac cine acquisition by various models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: kt-FOCUSS (ME/MC) - (31.3, 0.920), CRNN - (31.9, 0.921), ME-CRNN - (32.2, 0.926).

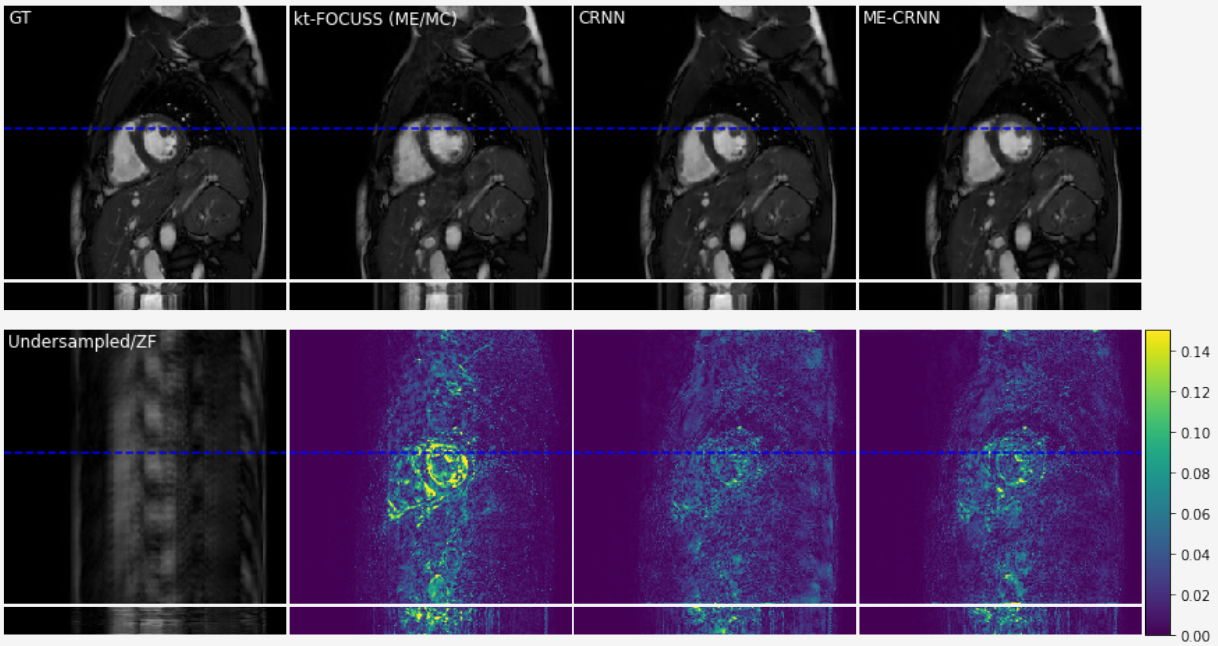

Figure 3.20: Comparison of the reconstruction of a x16 cardiac cine acquisition by various models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: kt-FOCUSS (ME/MC) - (34.1, 0.942), CRNN - (35.9, 0.949), ME-CRNN - (36.2, 0.956).

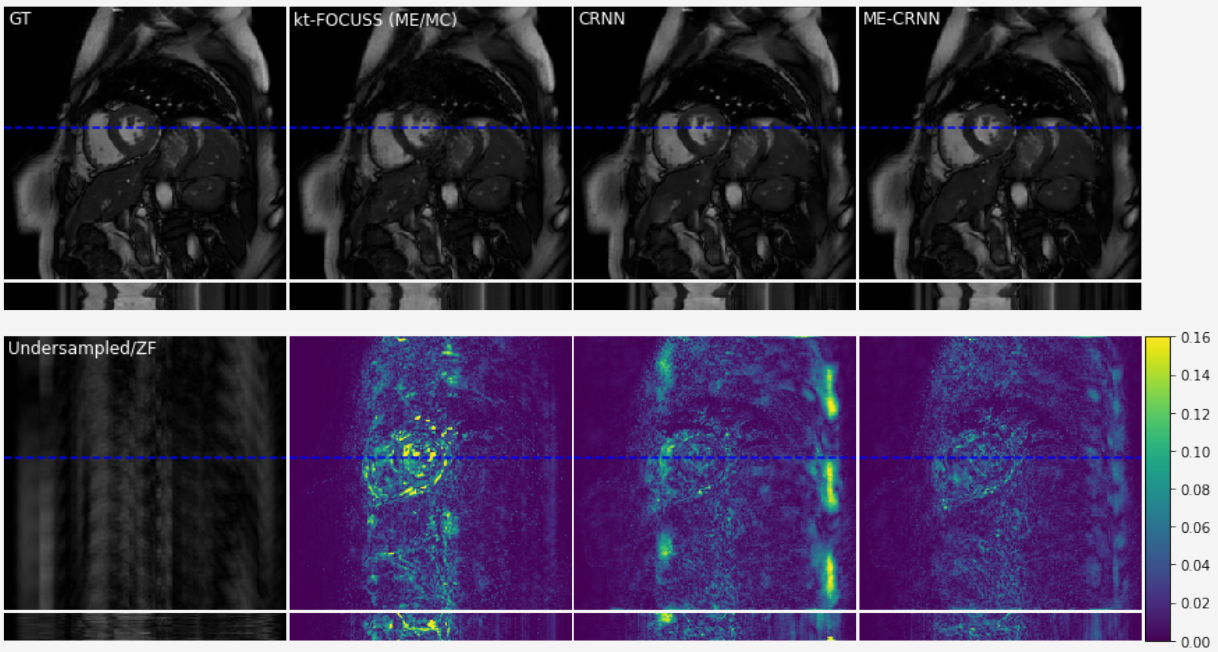

Figure 3.21: Comparison of the reconstruction of a x16 cardiac cine acquisition by various models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: kt-FOCUSS (ME/MC) - (33.7, 0.930), CRNN - (32.3, 0.913), ME-CRNN - (35.9, 0.942).

Table 3.3: Quantitative metrics for the performance of models on the [93] dataset

Model	PSNR	SSIM	ΔPSNR	ΔSSIM
CRNN	31.0 ± 1.6	0.891 ± 0.021	0.0 ± 0.0	0.000 ± 0.000
ME-CRNN	31.4 ± 1.9	0.897 ± 0.022	0.4 ± 0.7	0.006 ± 0.007
ME-CRNN (HQ MF)	$\mathbf{32.1 \pm 1.6}$	$\mathbf{0.907 \pm 0.017}$	$\mathbf{1.1 \pm 0.5}$	$\mathbf{0.015 \pm 0.007}$
2D ME-CRNN (60its)	27.0 ± 2.2	0.800 ± 0.051	-4.0 ± 1.0	-0.091 ± 0.031
2D ME-CRNN (60its, HQ MF)	29.0 ± 1.8	0.849 ± 0.035	-2.0 ± 0.7	-0.042 ± 0.017
2D BiCRNN	30.6 ± 1.8	0.879 ± 0.027	-0.4 ± 0.5	-0.012 ± 0.007
2D ME-BiCRNN	31.1 ± 2.3	0.891 ± 0.037	0.1 ± 1.0	-0.000 ± 0.019
2D ME-BiCRNN (HQ MF)	31.9 ± 2.1	0.902 ± 0.033	0.9 ± 0.9	0.011 ± 0.016
kt-FOCUSS	25.7 ± 2.0	0.751 ± 0.051	-5.3 ± 0.7	-0.140 ± 0.035
kt-FOCUSS (ME/MC)	27.7 ± 1.6	0.820 ± 0.027	-3.3 ± 0.5	-0.071 ± 0.011

over kt-FOCUSS.

The control experiment, CRNN, showed a decrease in performance compared with ME-CRNN. This is possibly attributed to two factors:

1. *Poor exploiting of data* There is no direct mechanism to exploit data in the temporal direction in a physical manner [141]
2. *Optimisation* ME-CRNN contains a physical model to guide to denoising process which may lead to a more robust optimisation. The incorporation of the crude/imperfect motion estimate helps better exploit the data available to the optimisation procedure.

3.7.3.1 Out of distribution data

Using the data collected from [93], we show that the ME-CRNN performs better than the CNTL. However, different hospitals will use different scanner settings which may shift the in vivo data to an area out of the training distribution. In order to test performance in such scenarios, the denoisers were also evaluated on some randomly chosen data from the UK BioBank. It

Table 3.4: Quantitative metrics for the performance of models on the cardiac cines from the UK BioBank study

Model	PSNR	SSIM	ΔPSNR	ΔSSIM
CRNN	35.4 ± 1.3	0.954 ± 0.010	0.0 ± 0.0	0.000 ± 0.000
ME-CRNN	36.7 ± 1.5	0.965 ± 0.010	1.4 ± 0.5	0.010 ± 0.004
ME-CRNN (HQ MF)	$\mathbf{37.5 \pm 1.4}$	$\mathbf{0.968 \pm 0.009}$	$\mathbf{2.1 \pm 0.4}$	$\mathbf{0.014 \pm 0.004}$
2D ME-CRNN (60its)	31.2 ± 1.6	0.915 ± 0.023	-4.2 ± 1.1	-0.039 ± 0.016
2D ME-CRNN (60its, HQ MF)	33.4 ± 1.4	0.941 ± 0.015	-1.9 ± 0.8	-0.014 ± 0.008
2D BiCRNN	34.4 ± 1.5	0.948 ± 0.012	-1.0 ± 0.6	-0.006 ± 0.005
2D ME-BiCRNN	36.5 ± 1.5	0.964 ± 0.010	1.2 ± 0.6	0.010 ± 0.004
2D ME-BiCRNN (HQ MF)	37.3 ± 1.4	$\mathbf{0.968 \pm 0.009}$	1.9 ± 0.6	$\mathbf{0.014 \pm 0.004}$
kt-FOCUSS	32.5 ± 1.8	0.927 ± 0.024	-2.9 ± 1.5	-0.027 ± 0.020
kt-FOCUSS (ME/MC)	34.3 ± 1.8	0.945 ± 0.018	-1.1 ± 1.4	-0.009 ± 0.014

should be noted that for the UK BioBank, only magnitude images were available and so we used synthetically generated phase maps to break any potential k-space symmetry [116] [3].

The expected behaviour is that supervised methods should show a drop in performance relative to unsupervised methods such as kt-FOCUSS. Whilst this is indeed the case, ME-CRNN still shows a significant advantage over CRNN. It should be noted that, whilst it appears that kt-FOCUSS (ME/MC) generally performs comparatively to the CRNN experiment, from Figures 3.19-3.21 it can be observed that the motion-rich region of interest (ROI) shows the significant motion corruption that generally exists in such algorithms.

3.7.3.2 Going beyond crude motion estimators

The high quality motion field, denoted in the results with 'HQ MF', were also used to generate reconstructions. The high quality motion fields were generated from a 5 cascade ME-CNN (see section 3.5). We note that whilst the crude motion estimate from a simple autoencoder-like

[3] Real valued functions contain conjugate symmetry in their Fourier domain. i.e. $f * (-x) = f(x)$.

network performed a satisfactory reconstruction that were better than the CRNN baseline, using a higher quality motion estimate resulted in far superior reconstructions, particularly in terms of PSNR.

3.7.4 Conclusion

By using a single pretrained motion estimator network, we were able to train a version of the ME-CNN which performs a denoising step after every 'DCMAC' step which involves a motion warp followed by data consistency. We also used a CRNN architecture to help alleviate the memory constraints of these large networks whilst also providing higher fidelity results [113]. Our experiments verified the use of motion in image reconstruction and show that using a single crude motion estimate is sufficient to provide reconstruction advantages.

For future work, it is necessary to train the motion estimator jointly with the ME-CRNN which was not investigated in this section. Furthermore, ways to extend the single motion estimator to a motion estimator within each recurrent unit should be explored in hope of achieving stable end-to-end training without gradient explosions and other potential training optimisation instabilities.

Chapter 4

Improving the ME-CNN

In this chapter, we build upon the work on motion-based MRI reconstruction from the previous chapter. In the first section, we focus on architectural improvements that are able to generate more refined motion estimates and better intermediate reconstructions that subsequently lead to high fidelty reconstructions. We compare this improved network, the ME-CNNv2, to various state-of-the-art reconstruction methods and find significant improvements in the case of aggressive acceleration rates. In the second section, we introduce a possible use case of abundant segmentation data from the UK BioBank study to aid in the reconstruction process. We find that the use of segmentation data is able to better regularise the motion estimator in the proposed motion-based reconstruction network. This ultimately leads to higher fidelity reconstructions.

4.1 ME-CNNv2

In this chapter, we introduce an improved version of the network from Chapter 3 called ME-CNNv2. The network from section 3.5 was reimplemented from TensorFlow 1.8 to TensorFlow 2.2 which resulted in improvements in memory efficiency and training time. This improvement allowed us to conduct studies into an enhanced architectural design of our proposed network. In particular, it allowed us to increase the N_{dcmac} iterations used in the network. The proposed architecture follows the decomposition of section (3.3) more closely by incorporating a number of changes:

1. *Larger number of unrolled iterations* In the original unrolled optimisation in section 3.3, an iteration consisted of a DCMAC step followed by a denoising step. In the ME-CNN implementation in section 3.5, the denoiser sees the output of the DCMAC step. However, it also sees other outputs generated by repeatedly applying the DCMAC step (without any intermediate denoising). The denoiser sees outputs with $N_{\text{ydcmac}} = \{1...30\}$ number of DCMAC steps which are referred to as y-DCMACs. It also receives the output from applying the DCMAC step $N_{\text{xdcmac}} = 60$ times by with the original k-space acquisition as input. In ME-CNNv2, we remove the latter as an input to the denoiser. Furthermore, we remove the y-DCMACs and replace them with a single cine generated by applying the DCMAC step K times to the output of the previous cascade.

   We performed a grid search for the possible values of K at each cascade, $c = \{1...N_c\}$ with the number of cascades $N_c = 3$. We denote this value at each cascade as K_c. We found that $K_c = 60 \forall c \in \{1...N_c\}$ gave the best reconstructions when setting identical weight initialisation and using a deterministic GPU setting which resulted in our contribution to a popular determinism-in-TensorFlow repository [135, 133]. This value of $K_c = 60$ was not possible with our previous implementation in TensorFlow 1.8. It should also be noted that this TensorFlow update allowed us to validate on values as large as $K_c = 360$, something which was not possible with the original ME-CNN implementation.

2. *Better Motion Estimator Network* Memory efficiency allowed us to increase the number

of feature maps in the downsampled U-net from 16 to 64 (see Figure 3.3). This allows for a higher detail motion estimate to be generated, improving the training stability and reconstruction fidelity at test time.

3. *More motion regularisation* We were able to add more motion terms to our loss function. In particular, we were able to include a loss on the optical flow warp from a frame to the ED and ES frames as well as a loss on the consecutive warp of a frame through the entire cine cycle in order to mitigate motion blurring artefacts in the image reconstruction.
4. *Intermediate losses* We were able to add weighted losses on the intermediate reconstructions from each cascade which helped guide the unrolled optimisation and particularly helped the motion estimators. In our previous implementation, adding these intermediate losses led to memory issues which meant other parts of the network had to be compromised.

4.1.1 MECNNv2 Loss Function

The MECNNv2 loss function can be decomposed into reconstruction and motion losses as shown in (4.1). With our notation, our motion field estimator, M_i^P, describes the warping of the frame i to a frame implied by the choice of P such that $P \in \{f, ED, ES\}$. If $P = f$, it warps frame i to the next sequential frame, $i + 1$. This is the main choice of p is the only type of warping used in forward inference. If $P = ED$, it warps i to the ED frame. If $P = ES$, it warps i to the ES frame. These latter two options are used for generalisation of the motion estimator but not required at forward inference. The ES and ED frames are chosen because they are the two extremes of the different cardiac phases. This allows us to better account for the full range of cardiac motion that might occur.

$$
\begin{aligned}
&L(\{m_1 ... m_{N_c}\}, \{M_1^{ED} ... M_{N_c}^{ED}\}, \{M_1^{ES} ... M_{N_c}^{ES}\}, \{M_1^f ... M_{N_c}^f\}; m_{gt}) = \\
&\quad \text{recon loss} + \text{main motion loss} + \text{additional motion generalisation losses} = \\
&\quad \sum_k^{N_c} \left[L_{\text{recon}}(m_k, m_{gt}) + L_{\text{f}}(\{M_{k,1}^f ... M_{k,T}^f\}, m_{gt}) + L_{g,k} \right]
\end{aligned}
\tag{4.1}
$$

The first part of the loss function is the weighted reconstruction loss:

$$L_{\text{recon}}(m_k, m_{gt}) = 2^{k-N_c}||m_k - m_{gt}||_2^2 \tag{4.2}$$

The second and most important part of the loss function involves warping each frame to the next frame immediately after itself along with spatial and temporal regularisation and additional regularisation term L_{rf}:

$$L_{\text{f}}(\{M_1^f ... M_T^f\}, m_{gt}) = \sum_i \left[\lambda||M_i^f m_{gt}^i - m_{gt}^{i+1}||_2^2 + c_5||\nabla^2 M_i^f||_2^2\right] + c_6 \sum_t ||\nabla_t^2 M^f||_2^2 + L_{rf}, \tag{4.3}$$

The term L_{rf} involves using the predicted transformation from frame i to $i+1$ (forward mode) to estimate the transformation from frame $i+1$ to i (reverse mode), without any further learned computation. This helps to ensure that the forward transformation used generates frames that can be reverted to their original state:

$$L_{\text{rf}}(\{M_1^f ... M_T^f\}, m_{gt}) = \sum_i \left[\lambda||\text{Inv}(M_i^f) m_{gt}^{i+1} - m_{gt}^i||_2^2\right] \tag{4.4}$$

It should be noted that $\text{Inv}(M)$ is a pseudo-inverse motion field which reverses the warp caused by M. This can be implemented as obtaining the individual horizontal and vertical velocity components that compose M, and individually warping them by the motion field M (i.e. itself). The negative of each individual component is then equal to the individual components of the pseudo-inverse.

4.1.1.1 Additional motion regularisation term, $L_{g,k}$

The third and final part of the loss function is composed of a series of additional motion loss terms that aid with generalisation of the motion estimator.

$$L_{g,k} = \Big(\sum_{P \in \{ED,ES\}} L_{\mathrm{P}}(\{M^P_{k,1} ... M^P_{k,T}\}, m_{gt}) \Big) + L_{\mathrm{s}}(\{M^s_{k,1} ... M^s_{k,T}\}, m_{gt}) \tag{4.5}$$

The first term in $L_{g,k}$ is a frame-to-ED loss function. This is shown in equation (4.6) and can be seen to encompass an optical flow term, spatial regularisation term and an ED-to-frame 'reverse' optical flow loss that ensures consistent optical flow fields are generated.

$$\begin{aligned} L_{\mathrm{ED}}(\{M^{ED}_1 ... M^{ED}_T\}, m_{gt}) = & \sum_i \Big[c_1 || M^{ED}_i m^i_{gt} - m^{ED}_{gt} ||^2_2 + c_2 || \nabla^2 M^{ED}_i ||^2_2 \Big] \\ & + c_3 \sum_t || \nabla^2_t M^{ED} ||^2_2 + L_{\mathrm{rED}}, \end{aligned} \tag{4.6}$$

$$L_{\mathrm{rED}}(\{M^{ED}_1 ... M^{ED}_T\}, m_{gt}) = \sum_i \Big[c_1 || \mathrm{Inv}(M^{ED}_i) m^{ED}_{gt} - m^i_{gt} ||^2_2 \Big] \tag{4.7}$$

The second term within the summation in $L_{g,k}$ is identical except uses the ES frame instead of the ED frame. The final part of $L_{g,k}$ is a term which warps the first frame of the cine to each of the other frames in the cine. This helps reduce drifting and blurring artefacts that may occur with optical flow in highly accelerated acquisitions and in particular in the DC-MAC set-up. Likewise, this is followed by a loss which warps the last frame of the cine to every other frame using the inverse of estimate motion field.

$$L_{\mathrm{s}}(\{M^f_1 ... M^f_T\}, m_{gt}) = \sum_i c_4 || M^f_i ... M^f_0 m^0_{gt} - m^i_{gt} ||^2_2 + L_{\mathrm{rs}}, \tag{4.8}$$

$$L_{\mathrm{rs}}(\{M^f_1 ... M^f_T\}, m_{gt}) = \sum_i c_4 || \mathrm{Inv}(M^f_i) ... \mathrm{Inv}(M^f)_{T-1} m^T_{gt} - m^i_{gt} ||^2_2, \tag{4.9}$$

The hyperparameters are chosen to be $c_1 = c_4 = 0.1\lambda$, $c_2 = 5e2c_1$, $c_3 = 5e3c_1$, $c_5 = 5e2\lambda$,

$c_6 = 5e3\lambda$, $\lambda = 50.0$.

4.1.2 Data

Along with the data from section 3.5 and the UK BioBank cardiac data, in this section, we also include an evaluation on a fetal cardiac dataset. This consists of 16 152×400 multi-slice volume cines, each with $T = 96$ frames, collected using 25 coils. These coils were linearly combined using the emulated single-coil (ESC) method to generate an estimate for a single-coil acquisition [175].

4.1.3 Experiments

We compare our proposal against some other suitable reconstruction candidates. We also use a motion compensated reconstruction algorithm known as kt-FOCUSS (ME/MC) which requires fully-sampled references. This allows for a better understanding of the strengths and limitations of our proposal in certain scenarios.

4.1.3.1 Non-denoised MECNN

As a demonstration of the capabilities of motion-based reconstruction, we choose to use the MECNNv2 but without the denoising CNN in each cascade. The motion estimator within each cascade sees the output of the previous cascade to better improve the motion estimate. However, the reconstruction output generated by each cascade is simply the x-DCMAC from section 3.5 where the DCMAC step is performed for 60 iterations but using the motion estimate of the current cascade.

4.1.3.2 Variational Network

The variational network is a learnable gradient descent based approach for MRI reconstruction. The optimisation of the reconstruction from section 3.1 takes places via gradient descent

but with a regularisation term inspired by the Field-of-Experts model [29]. This Field-of-Experts consists of learnable convolutional kernels (CNNs) and trainable activation functions (a weighted sum of Gaussians). The resulting optimisation resembles the following descent:

$$m_{i+1} = m_i - \sum_{j=1}^{N_c} (K_i^j)^T \Phi_i^j (K_i^j m_i) - \lambda_i E^*(Em_i - y), \tag{4.10}$$

where K_j^i are the convolutional kernels, λ_i is the step size and Φ_i^j is the trainable activation function.

4.1.3.3 kt-FOCUSS

In kt-FOCUSS, the aim is to exploit sparsity in the x-f domain of the underlying cine, m. We note that the notation for variable m is in the x-t domain. In the x-f domain, we denote our cine as ρ instead. The acquired k-space measurements in the k-t domain, y, is thus represented as:

$$y = \mathcal{F}\rho, \tag{4.11}$$

where the Fourier transform operates in the temporal domain also. Here, we provide some detail into kt-FOCUSS in order to demonstrate how a motion-based implementation of kt-FOCUSS works and thus better understand its limitations.

A typical solution for finding ρ is the minimum norm but this may imply too strong a constraint on the underlying baseline signal $\bar{\rho}$ such that $\rho = \bar{\rho} + \Delta\rho$, where $\Delta\rho$ is the residual signal. This may lead to energy smoothing and does not sparsify the solution for the underlying signal [35]. Instead, an L1-norm is used in combination with a weighting matrix, W, to appropriately sparsify $\Delta\rho$ changing the optimisation problem:

$$\min ||\Delta\rho||_2, \;\; s.t. ||y - \mathcal{F}W\Delta\rho||_2 \leq \epsilon, \tag{4.12}$$

where ϵ is the desired precision of the reconstruction. Using Lagrange multipliers, the constrained iterative solution can be found to be:

$$\rho_i = \bar{\rho} + \theta_n \mathcal{F}^H (\mathcal{F}\theta_n \mathcal{F}^H + \lambda \mathcal{I})^{-1}(y - \mathcal{F}\bar{\rho}), \tag{4.13}$$

where $\theta_n = W_n W_n^H$ (with W_n updated according to equation (4.14)) and λ is the Lagrange multiplier.

$$W_n = \begin{pmatrix} |\rho_{n-1}(1)|^p & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & |\rho_{n-1}(N)|^p \end{pmatrix}, 1/2 \leq p \leq 1 \tag{4.14}$$

In our experiments W is initialised in two different ways:

1. *Temporal Average* The data from all frames is combined to provide a better estimate of any particular frame. This is referred to as simply 'kt-FOCUSS' in our studies
2. *Motion Compensation* Using a fully sampled reference, a block matching algorithm is used to estimate a dense motion field that is used to relocate pixels in the fully sampled reference to any particular frame, creating a motion compensated frame [42]. The block matching algorithm requires an estimate of the target frame in which the above kt-FOCUSS (with temporal average) is used. The motion compensated frames are used in initialisation. This is referred to as 'kt-FOCUSS (ME/MC)' in our studies to denote motion estimation followed by motion compensation ending with kt-FOCUSS.

It should be noted that while kt-FOCUSS (ME/MC) provides an initialisation that provides very enhanced reconstructions, the requirement of fully-sampled reference frames imposes a limit to the possible acceleration rate which is not desirable. Furthermore, fully-sampled frames may not be easily acquired in certain clinical scenarios.

4.1.3.4 ME-CNN

This is the original ME-CNN implementation from section 3.5. The network is trained with the architectural and loss function updates as described in the MECNNv2 improvements. However, the DCMAC structure is slightly different. Each cascade receives an x-DCMAC of 60 iterations (in which the initial input is the zero filled reconstruction). Additionally, it receives 30 y-DCMACs varying in the number of DCMAC steps applied from 1 to 30, noting that the initial input is the output of the previous cascade. In this interpretation, the network sees all these varying iteration numbers and can choose the iterations which optimise the end reconstruction the best. However, the disadvantage compared with the MECNNv2 is that the maximum number of DCMAC steps ever used on the previous cascade is 30 compared to the MECNNv2 which uses 60.

4.1.3.5 DC-CNN

The DC-CNN here is the same as described in 3.5.

4.1.4 Results

We divide our results into three distinct parts in order to better understand the strengths and limitations of our proposed method:

1. Preliminary Investigation - The benefit of the components included in the ME-CNNv2 is studied against the ME-CNN and the ME-CNN without the denoising component. This serves as a sanity check.
2. Varying acceleration rates - We vary the acceleration rate from x2 to x51.2 in order to see if the proposed method fails to perform at certain undersampling factors
3. Full comparison - Here we compare the MECNNv2 against ME-CNN, variational network (VN), DC-CNN, kt-FOCUSS and kt-FOCUSS (ME/MC) including another dataset beyond the domain of the training dataset.

Table 4.1: Table of results for 3 cascade networks on x16 accelerated acquisitions. NB/ None of the networks are retrained on the UK BioBank Data and thus contain several out-of-training-domain examples.

Model	PSNR	SSIM	ΔPSNR	ΔSSIM
Our Data				
DCCNN	30.4 ± 2.4	0.890 ± 0.028	0.0 ± 0.0	0.000 ± 0.000
MECNN (no denoiser)	30.8 ± 2.2	0.885 ± 0.036	0.4 ± 1.6	-0.006 ± 0.031
kt-FOCUSS	30.5 ± 2.4	0.884 ± 0.036	0.1 ± 1.6	-0.006 ± 0.030
MECNN	31.1 ± 2.9	0.873 ± 0.050	0.7 ± 2.0	-0.017 ± 0.043
MECNN V2	$\mathbf{32.2 \pm 2.0}$	$\mathbf{0.900 \pm 0.029}$	$\mathbf{1.8 \pm 1.3}$	$\mathbf{0.010 \pm 0.023}$
BioBank				
DCCNN	32.3 ± 1.5	0.896 ± 0.013	0.0 ± 0.0	0.000 ± 0.000
MECNN (no denoiser)	32.7 ± 1.5	0.904 ± 0.017	0.4 ± 1.4	0.008 ± 0.015
kt-FOCUSS	$\mathbf{37.2 \pm 1.9}$	$\mathbf{0.955 \pm 0.015}$	$\mathbf{4.9 \pm 2.0}$	$\mathbf{0.059 \pm 0.017}$
MECNN	34.2 ± 2.5	0.921 ± 0.031	1.9 ± 2.4	0.024 ± 0.029
MECNN V2	35.1 ± 1.3	0.932 ± 0.011	2.8 ± 1.3	0.035 ± 0.012

4.1.4.1 Preliminary Investigation

In our preliminary investigation, we investigated 3-cascade models as a sanity check to understand the benefit of the individual components of the ME-CNNv2. For a fair comparison, we incorporated the motion network, motion loss/regularisation and intermediate loss changes in ME-CNNv2 into ME-CNN but not the changes in the DC-MACs seen by the denoiser block.

We also investigated a version of ME-CNNv2 which was trained without any denoisers. Instead, only the in-cascade motion estimators were trained. At every cascade, the motion estimate is derived from the DCMAC applied to the output of the previous cascade. In the case of the first cascade, the DCMAC is applied to the originally acquired k-space data (equivalent to x-DCMAC in section 3.5). The 'refined' motion estimate from the final cascade is then used to generate an x-DCMAC which is used as the reconstruction output. The results are shown in Table 4.1 and Figures 4.1-4.7.

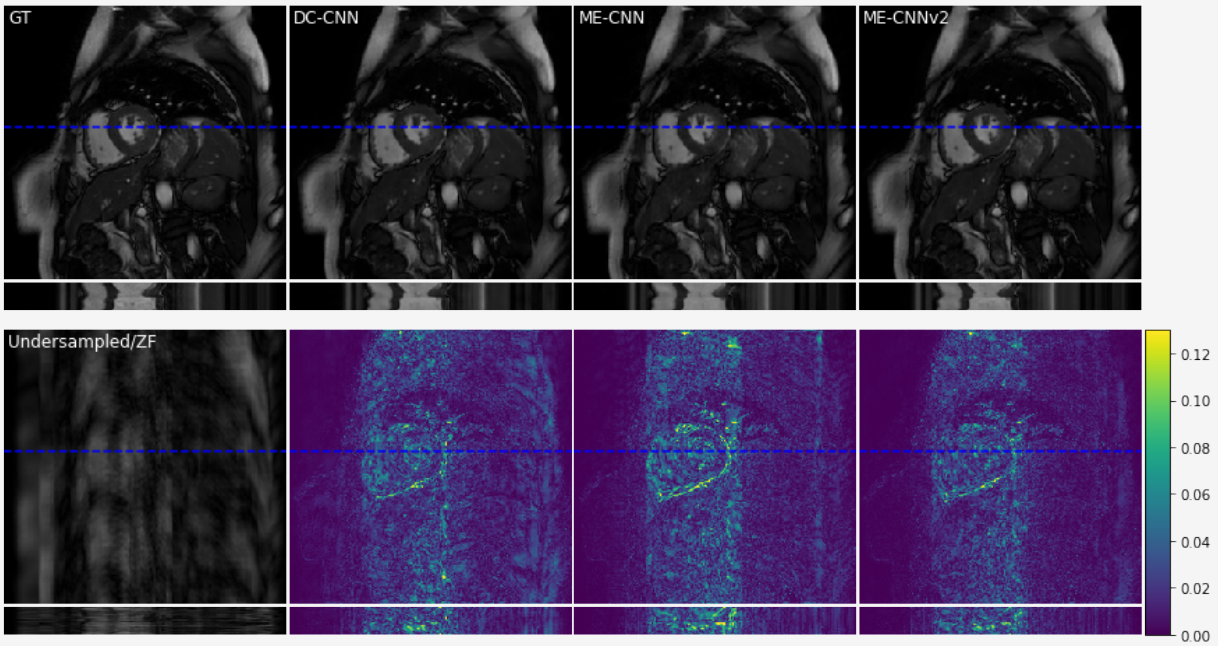

Figure 4.1: Comparison of the reconstruction of a x16 cardiac cine acquisition by various 3 cascade models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (36.4, 0.939), ME-CNN - (35.8, 0.935), ME-CNNv2 - (37.4, 0.957).

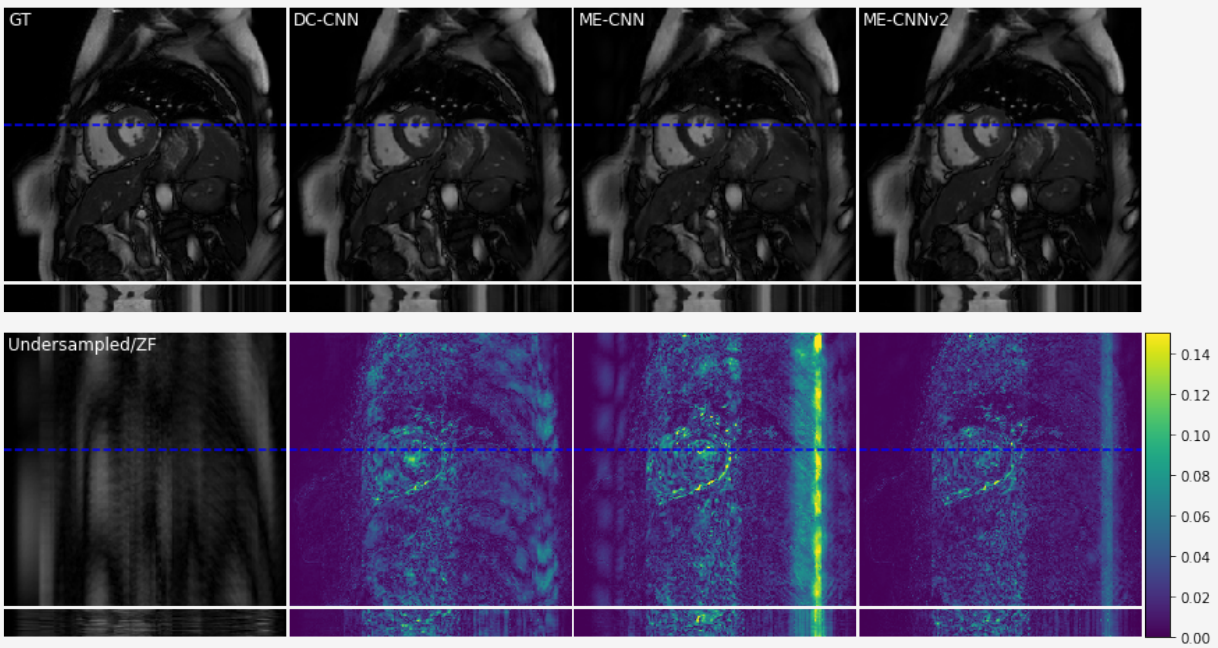

Figure 4.2: Comparison of the reconstruction of a x16 cardiac cine acquisition by various 3 cascade models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (34.8, 0.928), ME-CNN - (31.8, 0.869), ME-CNNv2 - (36.4, 0.947).

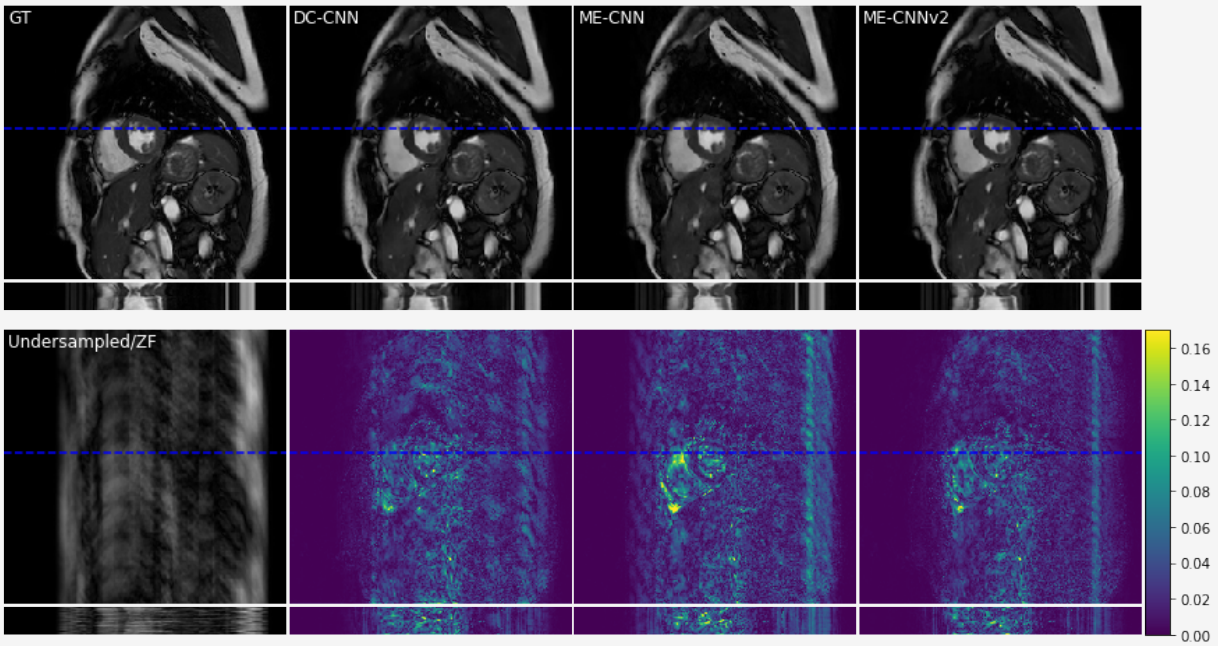

Figure 4.3: Comparison of the reconstruction of a x16 cardiac cine acquisition by various 3 cascade models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (33.7, 0.938), ME-CNN - (33.7, 0.923), ME-CNNv2 - (35.5, 0.957).

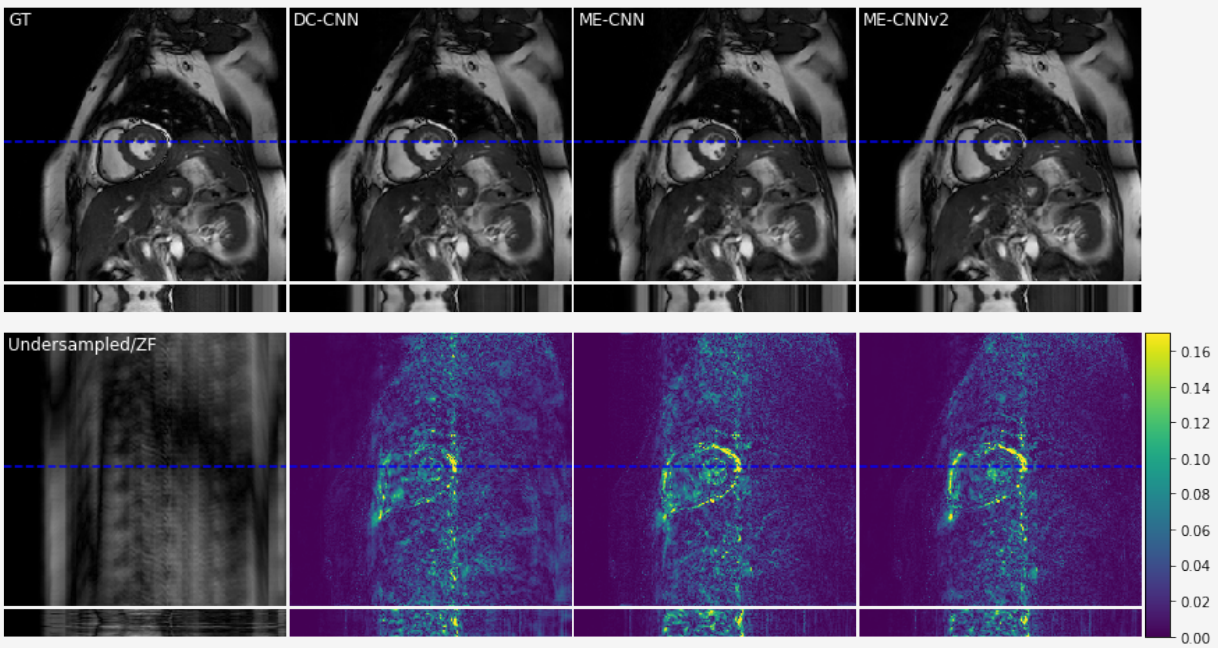

Figure 4.4: Comparison of the reconstruction of a x16 cardiac cine acquisition by various 3 cascade models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (32.9, 0.929), ME-CNN - (33.6, 0.931), ME-CNNv2 - (34.2, 0.944).

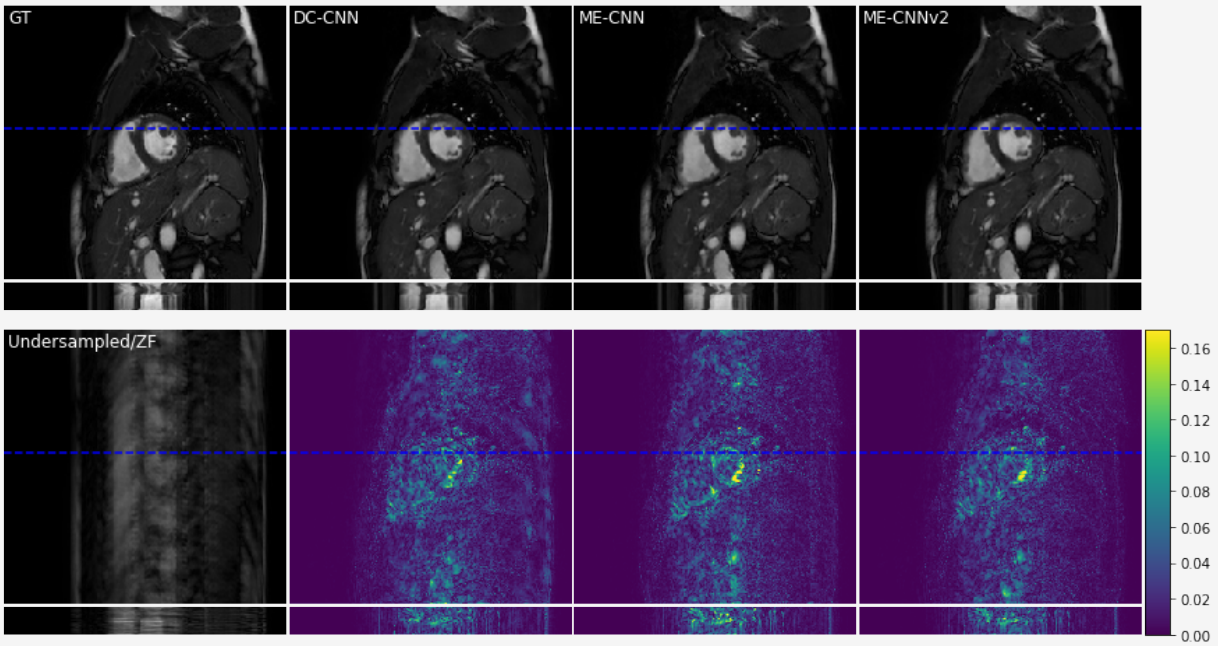

Figure 4.5: Comparison of the reconstruction of a x16 cardiac cine acquisition by various 3 cascade models. In this example, the advantage of MECNNv2 over DC-CNN is less pronounced. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (36.6, 0.954), ME-CNN - (35.6, 0.943), ME-CNNv2 - (37.1, 0.960).

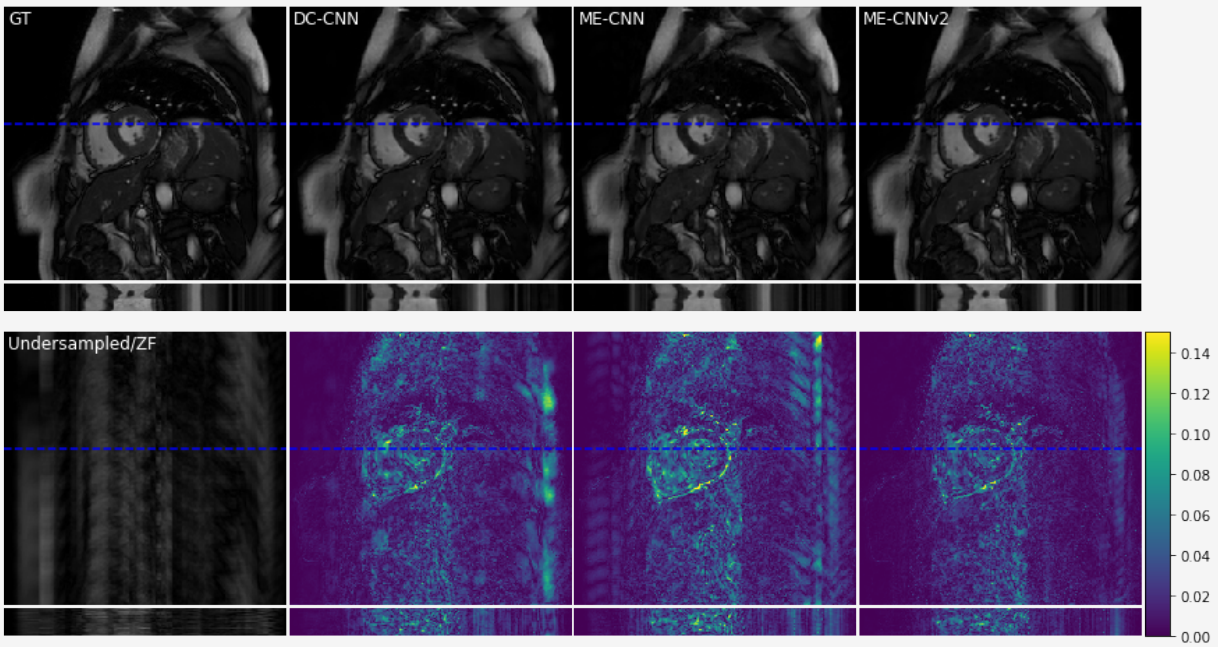

Figure 4.6: Comparison of the reconstruction of a x16 cardiac cine acquisition by various 3 cascade models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (34.9, 0.928), ME-CNN - (34.2, 0.903), ME-CNNv2 - (36.8, 0.954).

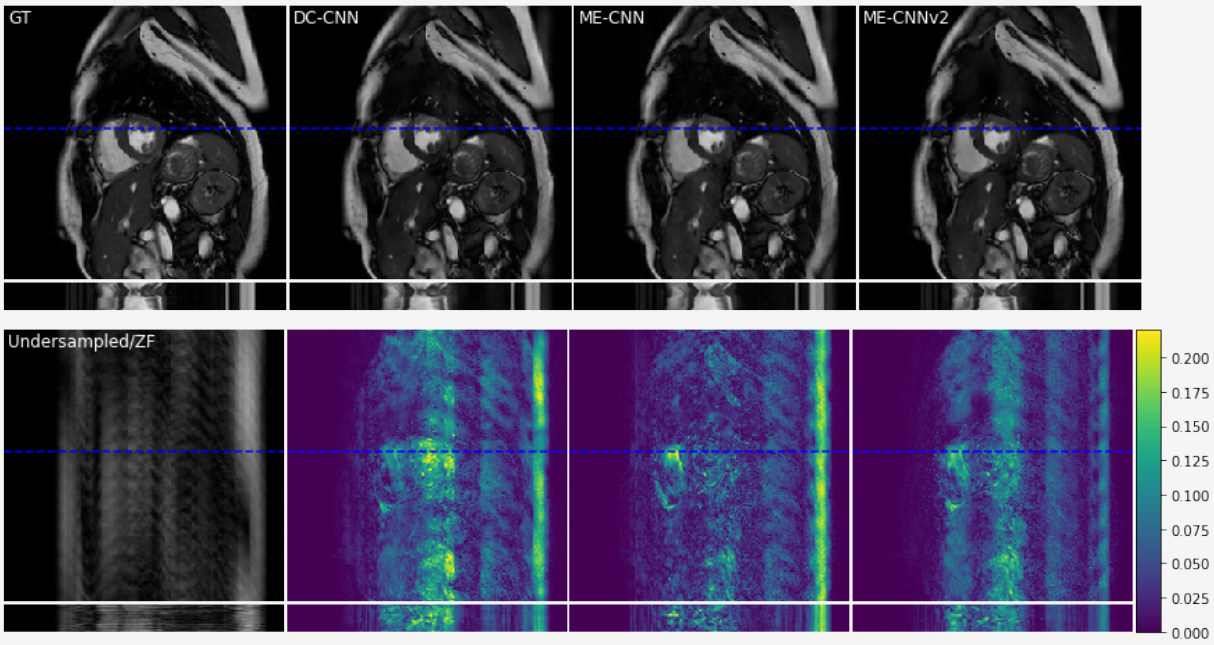

Figure 4.7: Comparison of the reconstruction of a x16 cardiac cine acquisition by various 3 cascade models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (27.5, 0.877), ME-CNN - (28.6, 0.889), ME-CNNv2 - (29.5, 0.905).

4.1.4.2 Varying acceleration rates

We investigate our the performance benefit of ME-CNNv2 changes compared to DC-CNN. The results of this are shown in Figure 4.8 and Figures 4.9-4.12.

4.1.4.3 Full comparison

In section, we compare against kt-FOCUSS (ME/MC) and variational network and introduce another dataset which are noisy, fetal images retrospectively undersampled, very different from the training domain. We also extend the number of cascades in our networks from 3 to 5 to show that this results improved performance when compared to kt-FOCUSS. The results of this are shown in Table 4.2 and Figures 4.13-4.16.

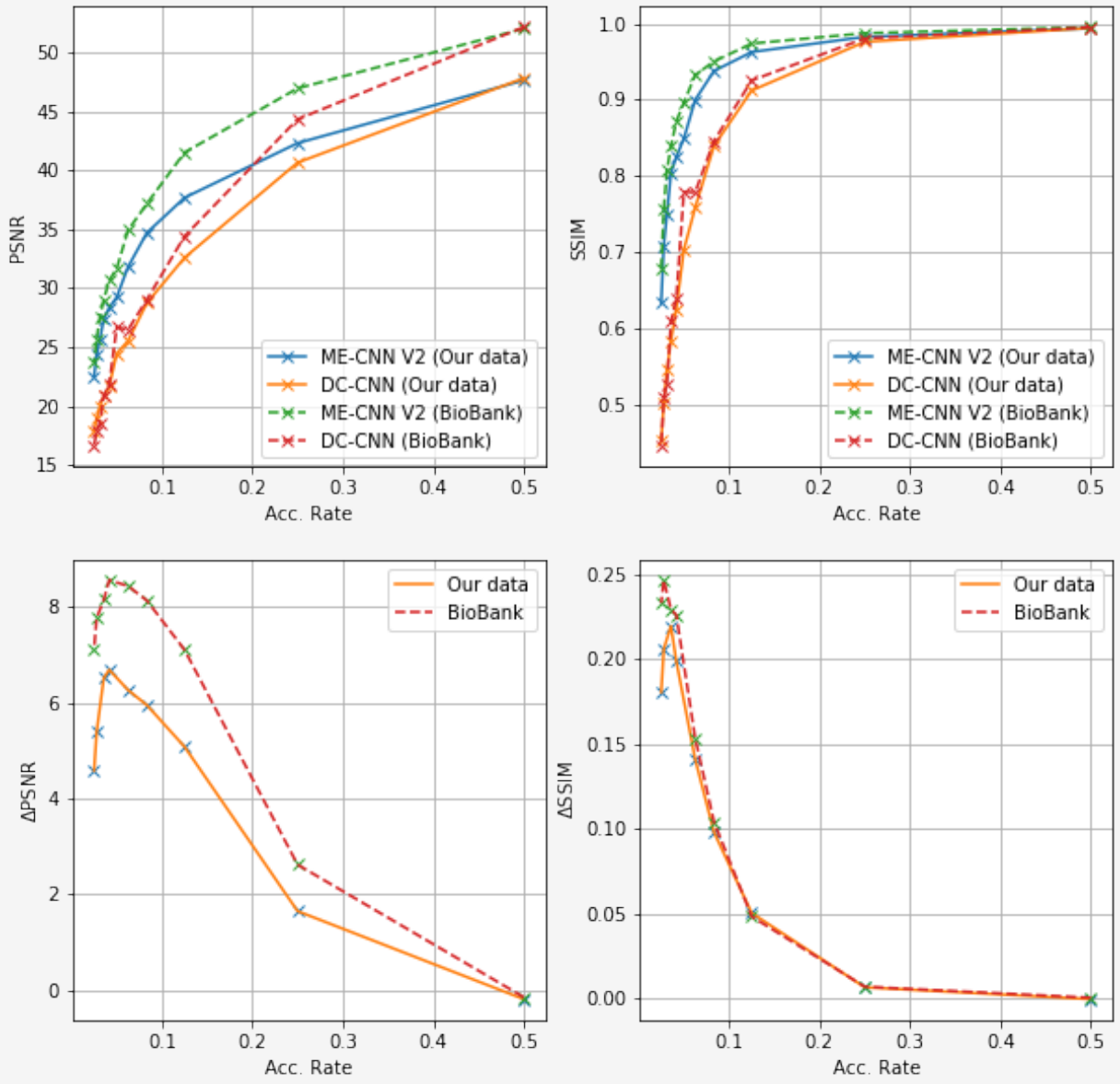

Figure 4.8: Results for a 3 cascade network on various acceleration rates. Note: Acceleration rate on the x-axis is plotted as the fraction of the k-space sampled.

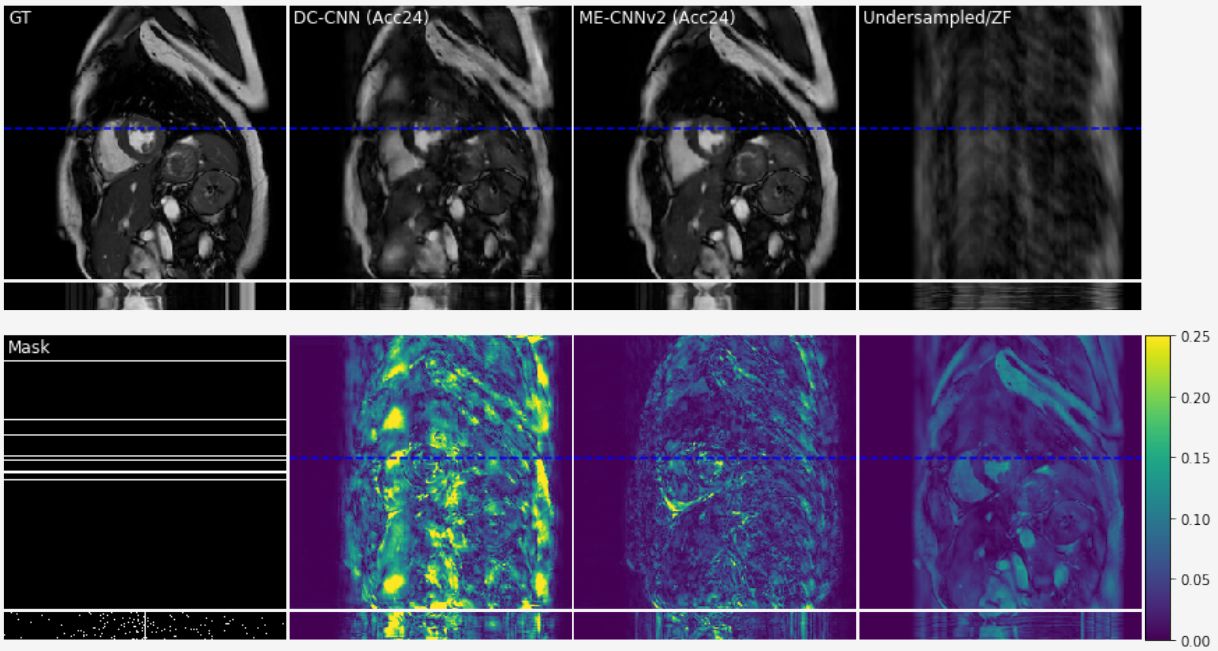

Figure 4.9: Comparison of the reconstruction of a x24 cardiac cine acquisition by two different reconstruction models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (21.4, 0.660), ME-CNNv2 - (28.1, 0.844).

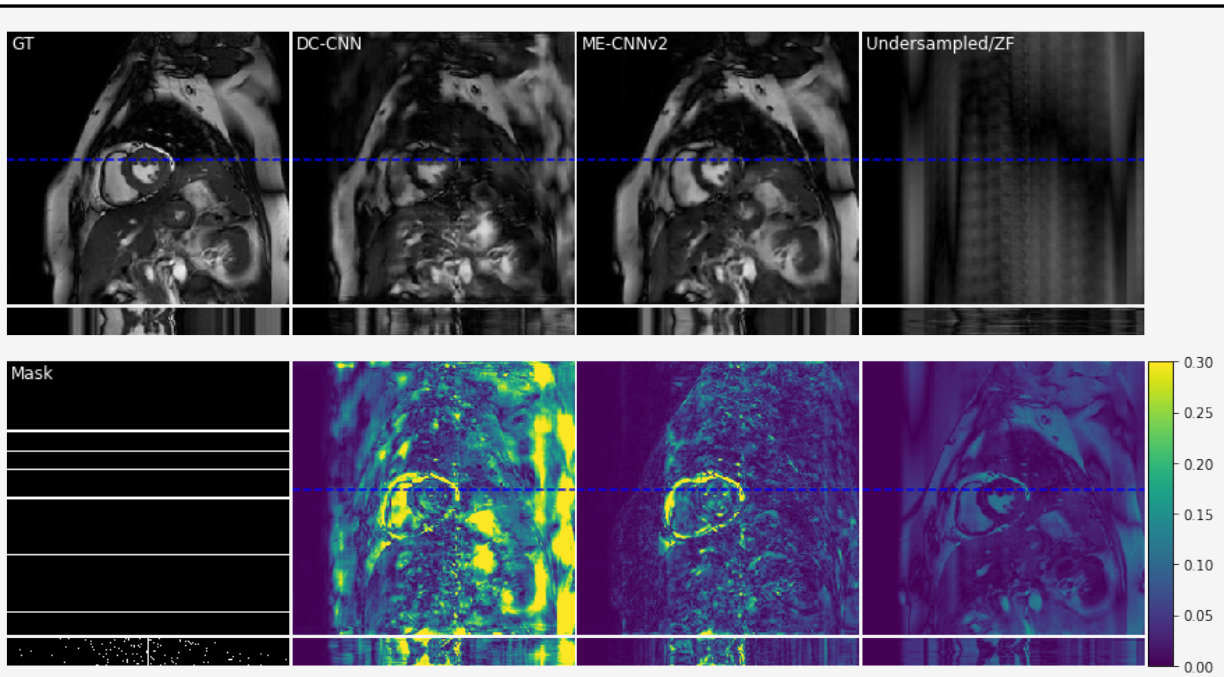

Figure 4.10: Comparison of the reconstruction of a x28 cardiac cine acquisition by two different reconstruction models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (18.8, 0.546), ME-CNNv2 - (26.2, 0.798).

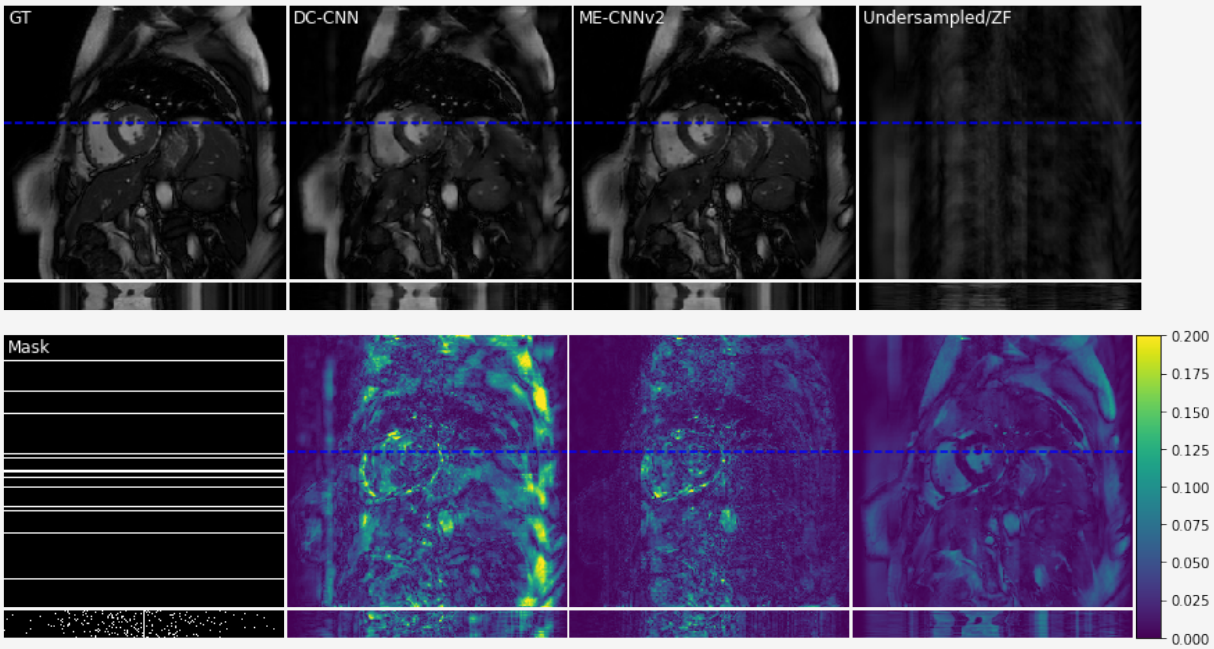

Figure 4.11: Comparison of the reconstruction of a x16 cardiac cine acquisition by two different reconstruction models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (27.5, 0.761), ME-CNNv2 - (32.8, 0.897). These reconstructions are from the same subject and sampling trajectory as those in Figures 4.14.

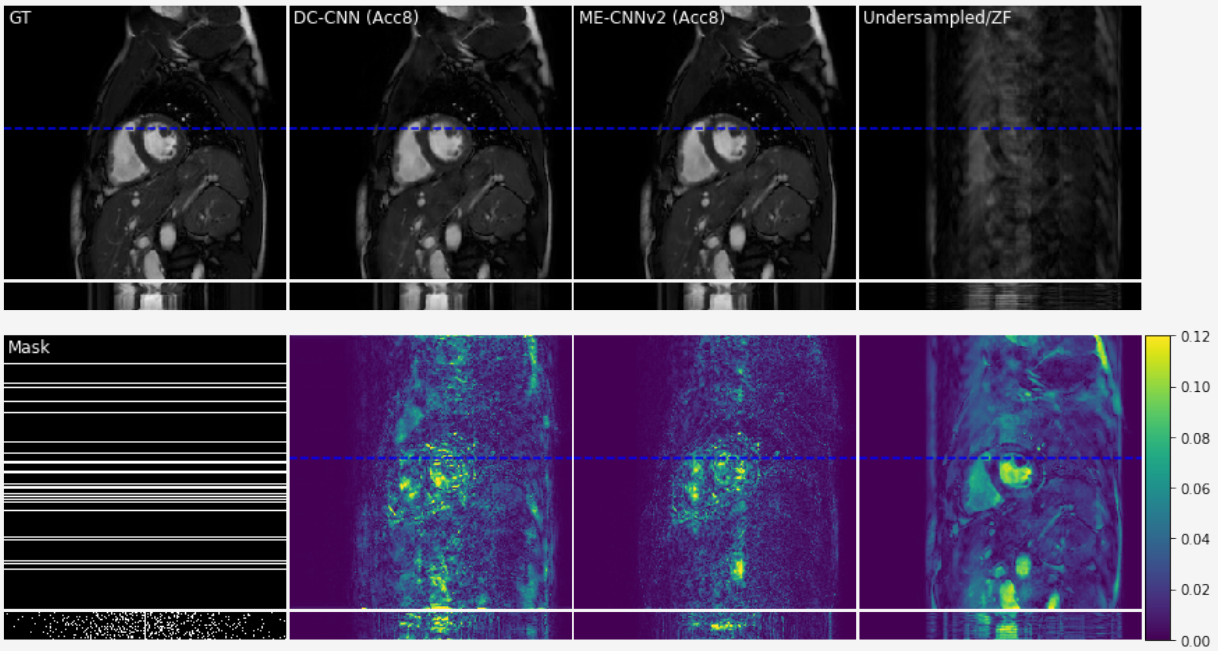

Figure 4.12: Comparison of the reconstruction of a x8 cardiac cine acquisition by two different reconstruction models. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (34.4, 0.929), ME-CNNv2 - (38.1, 0.963).

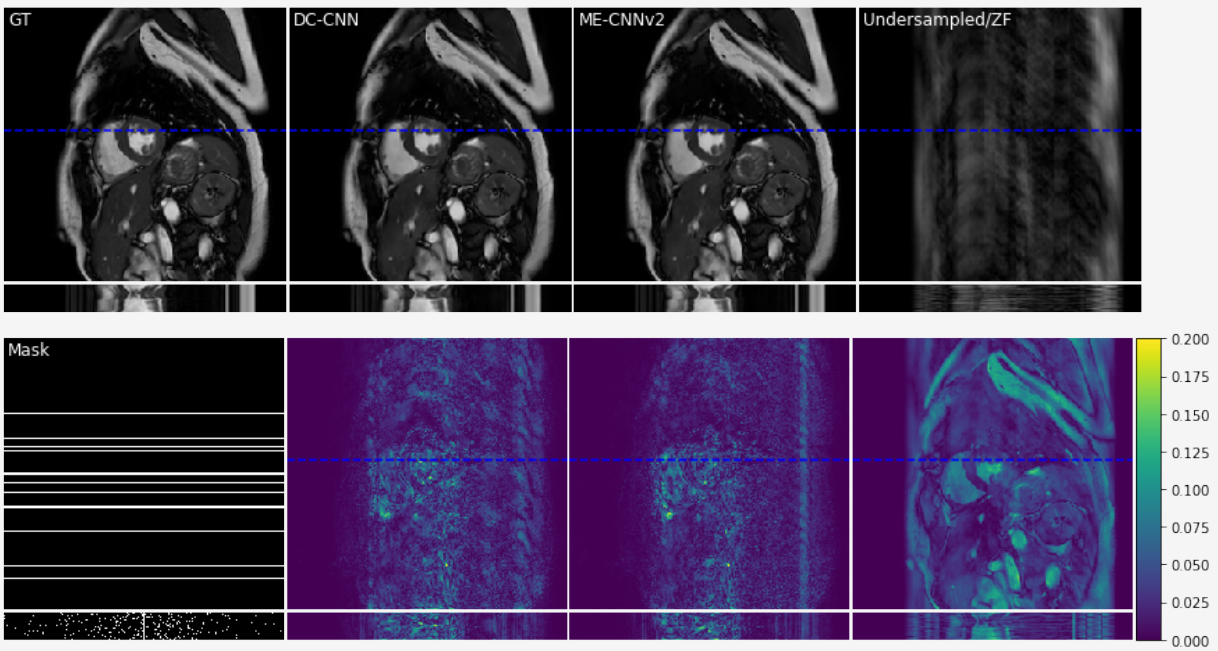

Figure 4.13: Comparison of the reconstruction of a x16 cardiac cine acquisition by two different reconstruction models with 5 cascades. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (33.7, 0.938), ME-CNNv2 - (35.5, 0.957).

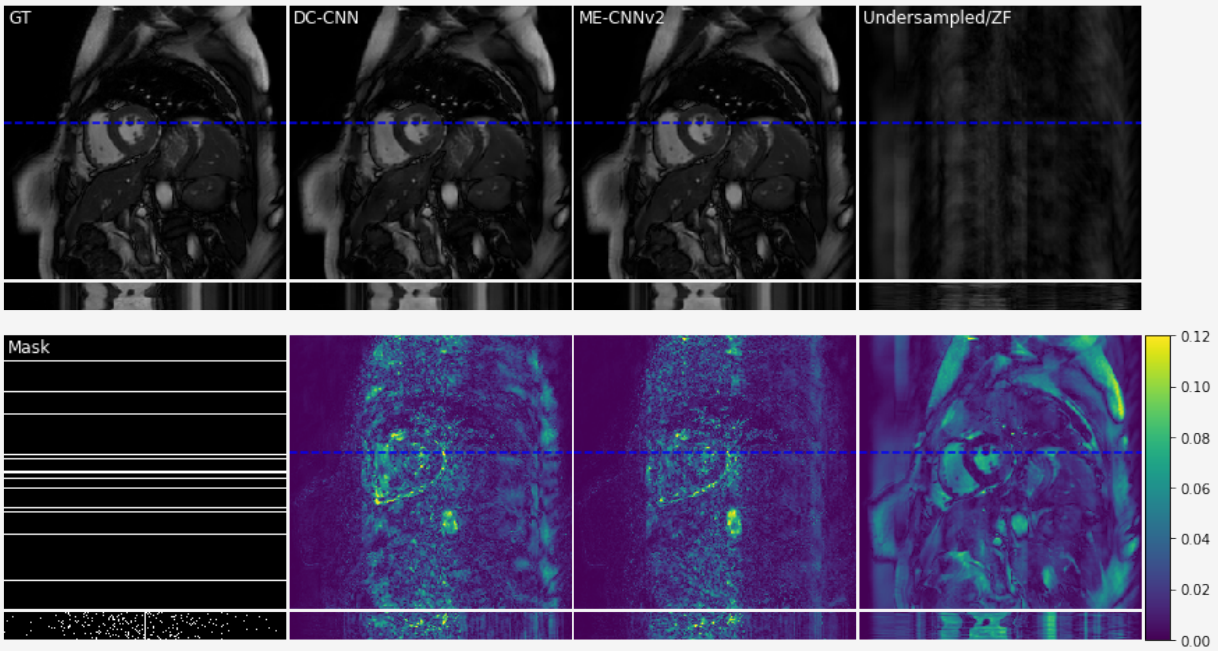

Figure 4.14: Comparison of the reconstruction of a x16 cardiac cine acquisition by two different reconstruction models with 5 cascades. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (35.4, 0.933), ME-CNNv2 - (37.0, 0.953).

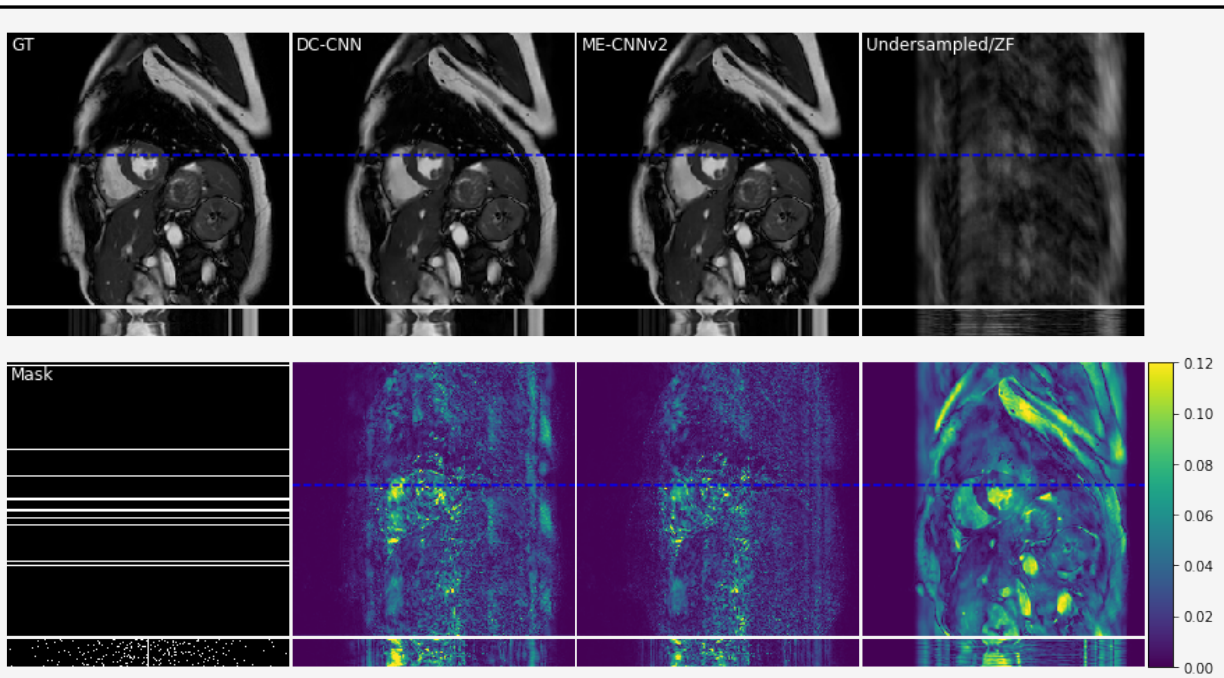

Figure 4.15: Comparison of the reconstruction of a x16 cardiac cine acquisition by two different reconstruction models with 5 cascades. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (33.1, 0.940), ME-CNNv2 - (35.7, 0.962).

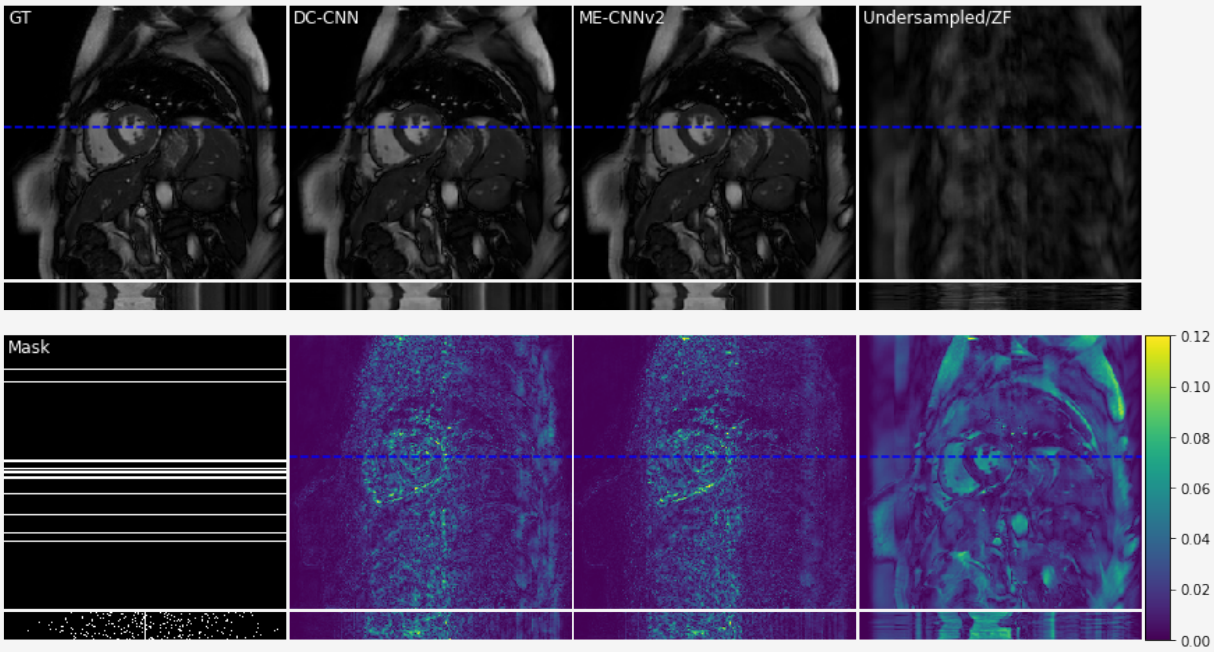

Figure 4.16: Comparison of the reconstruction of a x16 cardiac cine acquisition by two different reconstruction models with 5 cascades. The bottom row shows the difference with the ground truth. (PSNR, SSIM) for the reconstructions above: DC-CNN - (36.7, 0.946), ME-CNNv2 - (37.8, 0.958).

4.1.5 Discussion

4.1.5.1 Preliminary Investigation

The preliminary investigation shows that the ME-CNN without the CNN denoiser performs comparably to the DC-CNN (with denoisers). This highlights the potential power of motion exploitation in image reconstruction.

The L2 loss function used in this study is closely related to the PSNR metric used in evaluating the performance of these networks. Whilst the ME-CNN with and without the denoiser boast gains in PSNR compared to DC-CNN and kt-FOCUSS, the same does not hold true for SSIM for examples from the same dataset as the training domain. This highlights potential issues of loss functions for use in MRI reconstruction [140, 201, 83, 117, 125]. Interestingly, the DC-CNN performs comparably to the kt-FOCUSS method but when we shift the data domain outside of the training dataset, kt-FOCUSS (which is unsupervised), performs better than all other deep

Table 4.2: Table of results for 5 cascade networks on x16 accelerated acquisitions. NB/ Use of the BioBank data is not retrained and thus contains several out-of-training-domain examples. Italics for kt-FOCUSS (ME/MC) indicate that fully sampled reference frames are required and hence not a fair comparison.

Model	PSNR	SSIM	HFEN	VIF
Our Data				
VN	32.2 ± 2.1	0.894 ± 0.020	0.102 ± 0.019	0.815 ± 0.029
MECNNv2	$\mathbf{35.8 \pm 2.1}$	$\mathbf{0.952 \pm 0.013}$	$\mathbf{0.065 \pm 0.014}$	$\mathbf{0.899 \pm 0.022}$
MECNNv1	34.2 ± 1.8	0.923 ± 0.025	0.081 ± 0.021	0.866 ± 0.026
DC-CNN	34.9 ± 2.3	0.940 ± 0.017	0.074 ± 0.018	0.882 ± 0.028
kt-FOCUSS	30.5 ± 2.4	0.884 ± 0.036	0.130 ± 0.033	0.779 ± 0.043
kt-FOCUSS (ME/MC)	33.2 ± 1.7	0.934 ± 0.014	0.091 ± 0.014	0.840 ± 0.022
-	-	-	-	-
BioBank				
VN	35.1 ± 0.9	0.925 ± 0.008	0.068 ± 0.009	0.908 ± 0.016
MECNNv2	$\mathbf{39.9 \pm 1.4}$	$\mathbf{0.959 \pm 0.009}$	$\mathbf{0.038 \pm 0.007}$	$\mathbf{0.952 \pm 0.014}$
MECNNv1	38.0 ± 1.9	0.956 ± 0.017	0.048 ± 0.011	0.929 ± 0.018
DC-CNN	38.1 ± 1.6	0.952 ± 0.010	0.048 ± 0.009	0.940 ± 0.015
kt-FOCUSS	37.2 ± 1.9	0.955 ± 0.015	0.054 ± 0.013	0.919 ± 0.022
kt-FOCUSS (ME/MC)	39.6 ± 1.7	$\boldsymbol{0.978 \pm 0.006}$	0.040 ± 0.009	0.942 ± 0.018
-	-	-	-	-
Fetal				
VN	32.6 ± 2.0	0.831 ± 0.018	0.086 ± 0.021	0.761 ± 0.043
MECNNv2	36.5 ± 2.5	$\mathbf{0.932 \pm 0.022}$	0.048 ± 0.013	0.869 ± 0.039
MECNNv1	35.7 ± 2.5	0.910 ± 0.025	0.055 ± 0.015	0.848 ± 0.042
DC-CNN	34.9 ± 2.5	0.896 ± 0.031	0.060 ± 0.016	0.840 ± 0.043
kt-FOCUSS	$\mathbf{37.4 \pm 2.7}$	0.925 ± 0.026	$\mathbf{0.047 \pm 0.015}$	$\mathbf{0.881 \pm 0.047}$
kt-FOCUSS (ME/MC)	$\boldsymbol{38.3 \pm 2.9}$	$\boldsymbol{0.950 \pm 0.024}$	$\boldsymbol{0.042 \pm 0.015}$	$\boldsymbol{0.900 \pm 0.049}$

learning methods which is a common issue with deep learning approaches [81, 134].

Whilst shifting to the BioBank data leads to degradation in performance of deep learning methods compared to kt-FOCUSS, we find our proposed methods cope much better than the DC-CNN, with stark differences in PSNR and SSIM.

4.1.5.2 Varying acceleration rates

We also varied the acceleration rate from x2 to x51.2 (undersample factors from 0.02 to 0.5 as shown in Figure 4.8). This shows how the benefit of ME-CNNv2 and thus motion exploitation diminishes with more k-space sampling. This is expected as more temporally neighbouring data becomes available, the CNN kernel is able to better incorporate the data in the dealiasing process. Interestingly, even at relatively mild acceleration factors of x4 (0.25 undersampling), there is still a PSNR and SSIM advantage of MECNNv2 against DC-CNN. At x2 acceleration (0.5 undersampling), the advantage diminishes to virtually zero.

4.1.5.3 Full comparison

Here, we introduce comparisons against VN and kt-FOCUSS with motion estimation and compensation (ME/MC). We also maximise the number of cascades we can fit into GPU memory for ME-CNNv2 taking us the 5 cascades. Interestingly, by introducing two extra cascades, this is more than enough to alleviate the performance degradation (compared to kt-FOCUSS) suffered upon shifting to the BioBank dataset. In particular, ME-CNNv2 has significantly improved HFEN, VIF and PSNR and comparable SSIM compared to kt-FOCUSS. This is expected as in section 3.6, we demonstrated some aspects of robustness of motion based reconstruction in sections 3.5 and 3.6.

The BioBank dataset are still adult cardiac images and thus closer to the training dataset domain than the fetal cardiac cines. Whilst the ME-CNNv2 performs worse the kt-FOCUSS (ME/MC), it still performs drastically better than DC-CNN.

4.1.6 Conclusion

In this study we found that our improvements in the MECNNv2 have led to a robust approach to MR reconstruction in high acceleration settings with advantages over DC-CNN in shifts in data domain which may unexpectedly occur in a clinical setting. Further investigations of

this work should include the use of higher detail motion estimation networks and fitting more cascades into GPU memory. There is the possibility of extending this to volume acquisitions with the motion estimation occurring in 3D rather than just in-plane.

4.2 MSE-CNN: Joint Motion Estimation, Segmentation and Reconstruction

The ME-CNN architecture relies on motion estimation in order to complete the DCMAC step in equation (3.3). The motion estimator network is regularised by a Huber loss which ensures that motion estimates are smooth and reduces overfitting on noisy examples during the training process (especially with patch based training).

A previous study [114] explores the advantages of incorporating segmentation information and motion information into the same pipeline via a shared encoding process. The result was improved motion estimates as well as segmentations.

The ME-CNN relies on high quality motion estimates in the reconstruction process to ensure that post-warp cine prediction (the 'MAC' part of the DC-MAC) matches the subsequent data consistency step that is applied. Although section 3.7 shows that crude motion estimates aid in the reconstruction process, it also shows that higher quality motion estimate lead to drastically improved reconstructions.

If highly abundant segmentation data could be incorporated into the motion estimation part of the ME-CNN, it may lead to better motion estimates as in [114]. One could rationalise that a higher quality motion estimate would lead to better reconstructions thus connecting the reconstruction pipeline from segmentation to reconstruction.

In this section, we proposed a pipeline in which highly abundant UK BioBank segmentation date is used to regularise ME-CNN motion estimators based on the ideas of [114] that aim to improve motion estimation fidelity. We name our proposed approach the Motion-Segmentation Exploiting Convolution Neural Network (MSE-CNN).

4.2.1 Previous Work

In [59], GMM segmentations are used in a dictionary learning-based reconstruction that ultimately leads to sharper edges in the reconstruction of brain and cardiac images. The subsequent segmentations degrade in quality with acceleration rate but far less dramatically than having separate reconstruction and segmentation pipelines. [127] use neural networks to parameterise the components in the FISTA decomposition allowing end to end training to perform image reconstruction on cardiac data:

$$m_k = z_k - \text{FCN}\{\text{CNN}(z_{n-1}) + \text{FCN}(z_{n-1} - v_k) + \mathcal{F}^T(D\mathcal{F}v_k - y)\}, \tag{4.15}$$

where FCN is a fully connected network, v_k is a linear combination of m_{k-1} and m_{k-2} and z_{n-1} is the previous output from this same equation but without increment to k (instead increment n) and thus no update to the term v_k. They then add a U-net segmentor to the end of the pipeline and jointly optimise the segmentation and reconstruction to produce their proposed 'FR-Net'. This was evaluated on cardiac data with segmentations of the myocardium. Please see [127] for more details. [144] take a slightly different approach that integrates the segmentation representation into the reconstruction pipeline. In their study, a cascade of U-nets is used for reconstruction and at each cascade, the segmentor is the decoder of a U-net that reuses the encoder output of the reconstruction U-net. By having the segmentation and reconstruction share the same U-net encoder, trained jointly end-to-end, there is a boost in reconstruction and segmentation performance demonstrated on 3T T1 MRI brain data. It should be noted that the above approaches are either designed for single images or do not have a mechanism to exploit the temporal redundancy in cine MRI such as in ME-CNN.

4.2.2 Experimental Method

Inspired by [114], incorporating segmentation information into the motion estimation stage of the ME-CNN allows for a more detailed motion estimate for the subsequent reconstruction

stage of the ME-CNN. We do not require segmentations to be available to perform test-time reconstruction since segmentations are usually performed after post-processing. Instead, segmentation data is only used at training time to regularise the motion estimator both in terms of architectural design and in the optimisation landscape (via the addition of particular losses).

The two ways segmentations are used to regularise the motion estimator are as follows:

1. *Motion-Segmentation Encoded U-net with added segmentation loss* - By encoding segmentation information into the motion estimator, a better, more semantic representation can be learned that will result in both better motion estimates and segmentations as shown in [114].
2. *Dice loss for non-ED/non-ES segmentation predictions warped to ES/ED frames where radiologist/clinical segmentation is known* - This ensures that the motion field learned via the optical flow loss is consistent with warping predicted segmentations from one frame to another and thus provide a direct way in which segmentation information is used to regularise the motion estimator.

Producing segmentations as well as a motion estimate requires a lot of memory and is computationally intensive to train. In order to alleviate this issue, we use a single motion estimator and segmentation network that is shared by all cascades in the network. The input to the motion estimator-segmentation encoder is the undersampled/zero-filled reconstruction. We train our network end-to-end with the ME-CNNv2 loss function with two additional losses for the segmentations and warped segmentations:

$$\begin{aligned} L_{\text{MSE-CNN}} = L_{\text{ME-CNNv2}}(m_{gt}, y; \lambda) &+ \alpha_1 \text{Dice}(s_{\text{gt}}, s_{\text{pred}}) \\ &+ \alpha_2(\text{Dice}(s_{\text{pred,ED}}, M_{\text{ED}}(s_{\text{pred}})) \\ &+ \alpha_2 \text{Dice}(s_{\text{pred,ES}}, M_{\text{ES}}(s_{\text{pred}}))), \end{aligned} \tag{4.16}$$

where λ is the motion hyperparameter of the MECNNv2, s are segmentations, M represents a warp with the learned motion estimate to the ED (or ES) frame and α_1 and α_2 are hyperpa-

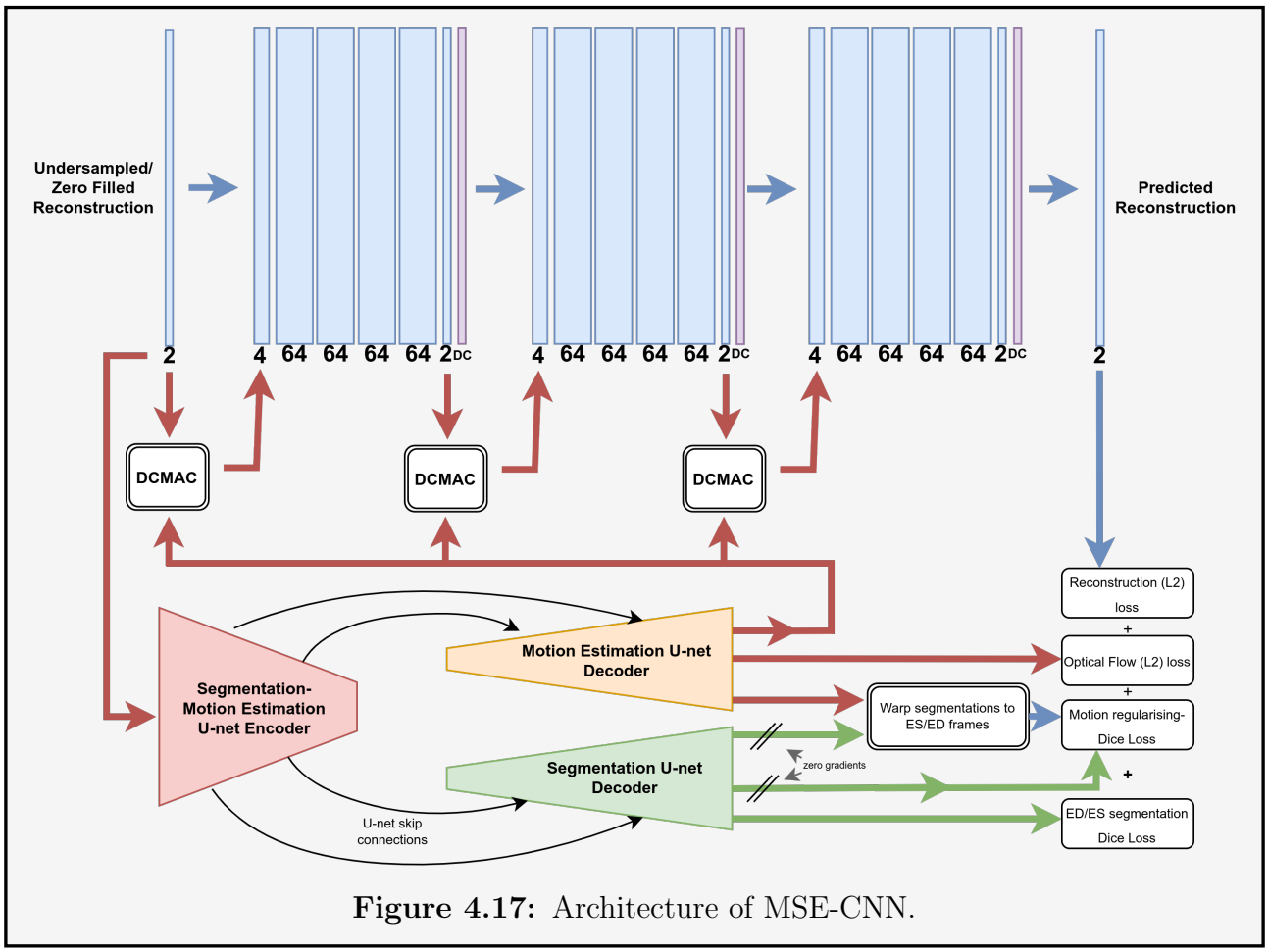

Figure 4.17: Architecture of MSE-CNN.

rameters. We chose $\alpha_1 = 1.0$ and $\alpha_2 = 0.2$.

We summarise the formulation of our method in Figure 4.17. It should be noted that the control experiment is identical except there is no segmentation part to the network and hence no segmentation or warped segmentation loss.

4.2.3 Results

The control experiment is used to determine whether the proposed method to introduce segmentation into the reconstruction pipeline creates enhanced reconstruction. The results of this comparison are shown in table 4.3. Some example reconstructions are shown in Figures 4.18-4.22.

Although the segmentation quality is not the aim of this study, we report obtained Dice scores.

Table 4.3: Results of MSE-CNN and a suitable control experiment on cardiac image and segmentation data from the UK BioBank study with synthetic phases to break k-space symmetry. DSC is the Dice score. Difference in metrics between the two methods is showed by the Δ.

Experiment	PSNR	SSIM	HFEN	VIF	RNMSE
CTNL	30.67 ± 0.79	0.918 ± 0.009	0.0502 ± 0.0052	0.825 ± 0.019	0.0756 ± 0.0089
MSE-CNN	$\mathbf{31.28 \pm 0.77}$	$\mathbf{0.929 \pm 0.009}$	$\mathbf{0.0460 \pm 0.0049}$	$\mathbf{0.842 \pm 0.018}$	$\mathbf{0.0704 \pm 0.0083}$
Δ	0.61 ± 0.11	0.011 ± 0.002	-0.0042 ± 0.0009	0.017 ± 0.003	-0.0051 ± 0.0012

The Dice score for the ED frame was found to be 0.787±0.036 and for the ES frame 0.741±0.046. Some example segmentations are also shown in this section.

4.2.4 Discussion

The quantitative results shown in table 4.3 clearly shown that the proposed method enhanced reconstruction quality across the entire cine with a signed Wilcoxon rank test $p \ll 0.01$ on the ROI across all metrics (for both rejecting null hypothesis that the methods are the same and accepting alternative hypothesis of enhanced performance). The results also show the difference images which highlight the MSE-CNN enhancement more clearly. In particular, the MSE-CNN generally produces a sharper image handling smaller details more appropriately.

4.2.5 Conclusion

In conclusion, we have demonstrated the advantage of segmentations in training motion-based reconstruction networks known as ME-CNNs. Whilst the MSE-CNN was shown to perform better than the variant of ME-CNN used as a control experiment, future work should include extending the motion estimation/segmentation to be unique to each cascade to ensure an optimal reconstruction is generated. Currently, the encoder never receives intermediate reconstructions from individual cascades which limits the possible optimisation of the motion

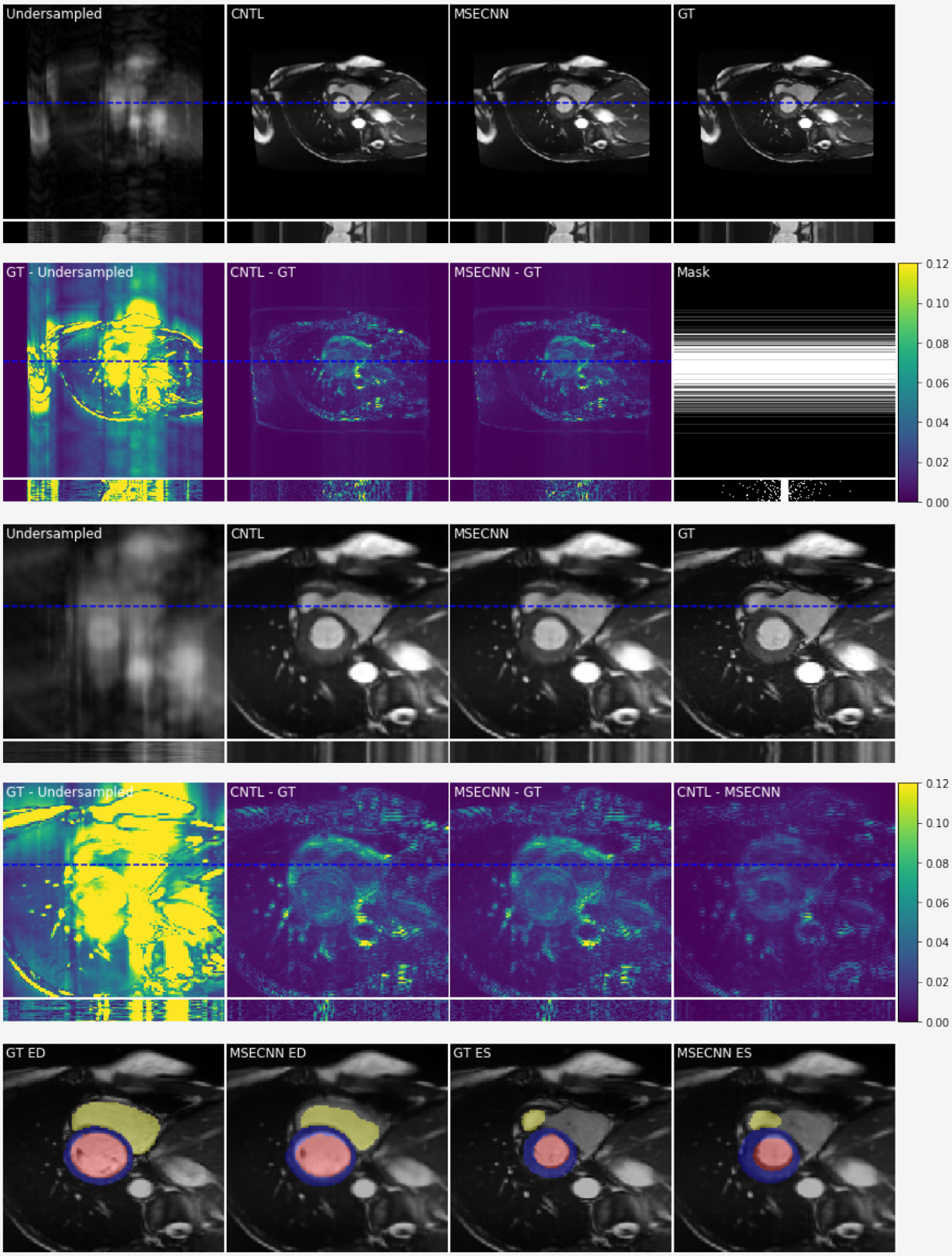

Figure 4.18: Comparison of the reconstruction of a x16 cardiac cine acquisition by the two different reconstruction models. The first and third row are the reconstructed images and enlarged central crop of said images. The second and fourth rows are the difference images with the row above. The last row are the segmentations generated by the MSE-CNN. (PSNR, SSIM) for the reconstructions above: CNTL - (36.3, 0.949), MSE-CNN - (37.0, 0.956). The Dice score of the (ED, ES) segmentations from the MSE-CNN were (0.843, 0.727).

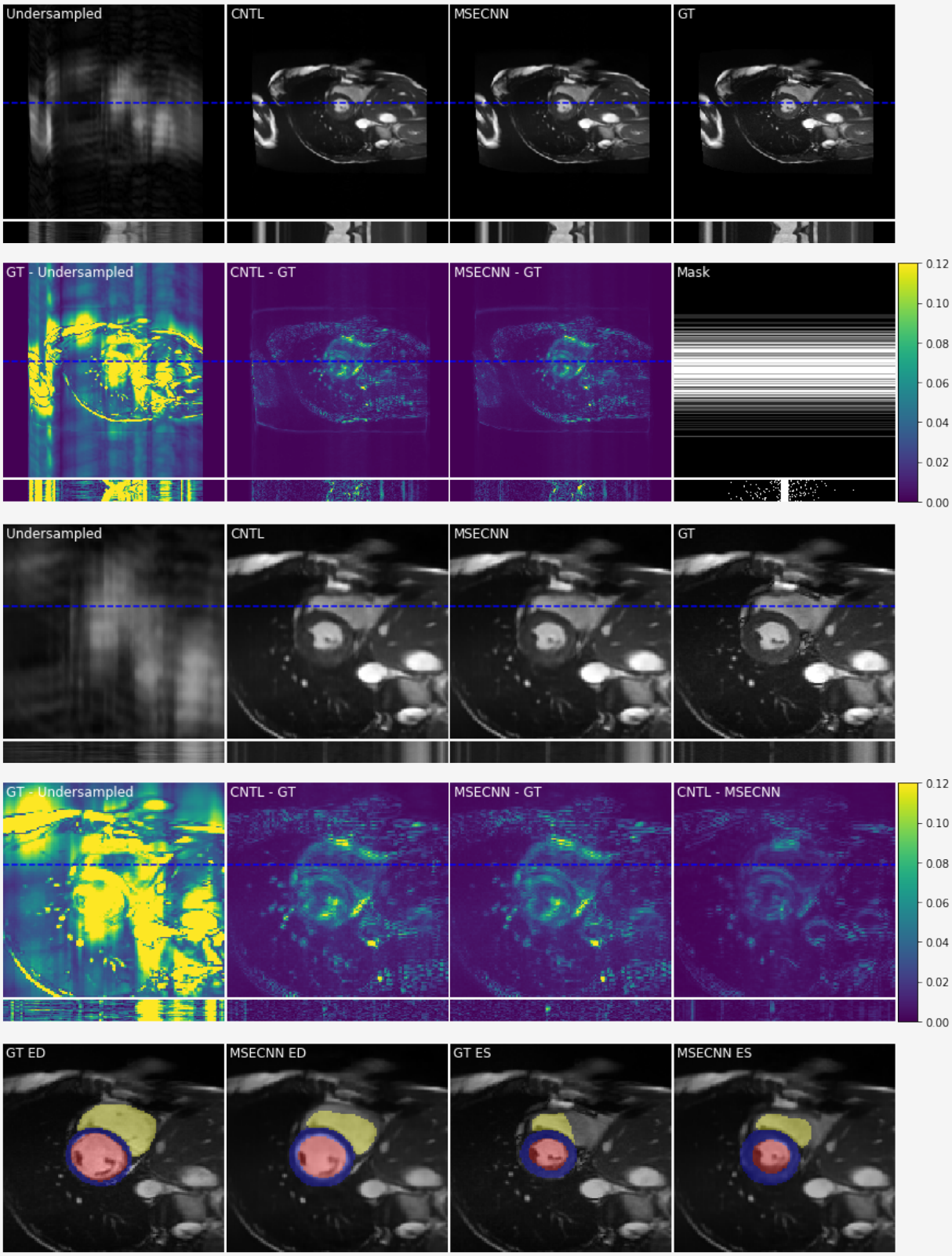

Figure 4.19: Comparison of the reconstruction of a x16 cardiac cine acquisition by the two different reconstruction models. The first and third row are the reconstructed images and enlarged central crop of said images. The second and fourth rows are the difference images with the row above. The last row are the segmentations generated by the MSE-CNN. (PSNR, SSIM) for the reconstructions above: CNTL - (36.8, 0.950), MSE-CNN - (37.3, 0.958). The Dice score of the (ED, ES) segmentations from the MSE-CNN were (0.803, 0.769).

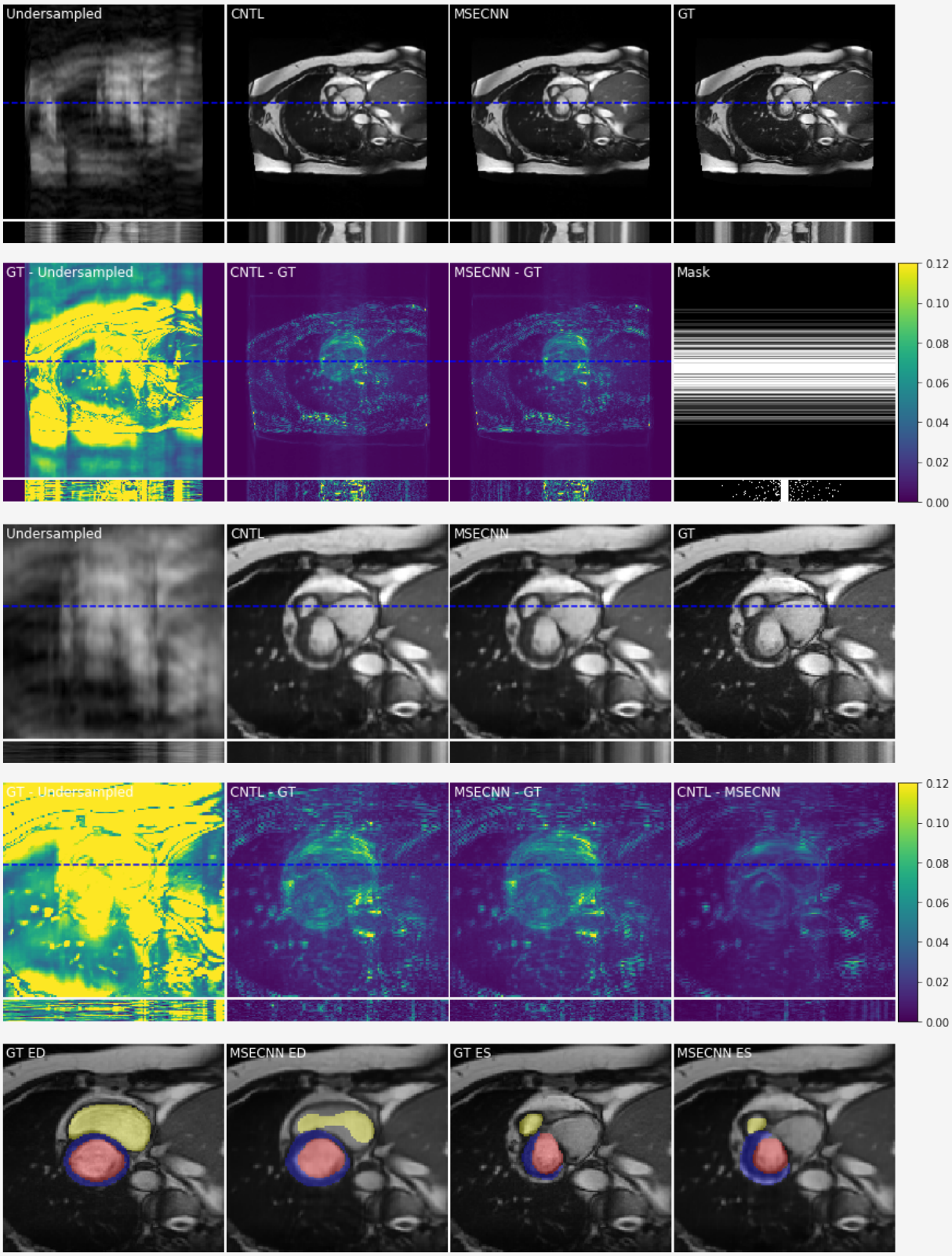

Figure 4.20: Comparison of the reconstruction of a x16 cardiac cine acquisition by the two different reconstruction models. The first and third row are the reconstructed images and enlarged central crop of said images. The second and fourth rows are the difference images with the row above. The last row are the segmentations generated by the MSE-CNN. (PSNR, SSIM) for the reconstructions above: CNTL - (34.6, 0.936), MSE-CNN - (35.0, 0.945). The Dice score of the (ED, ES) segmentations from the MSE-CNN were (0.717, 0.697).

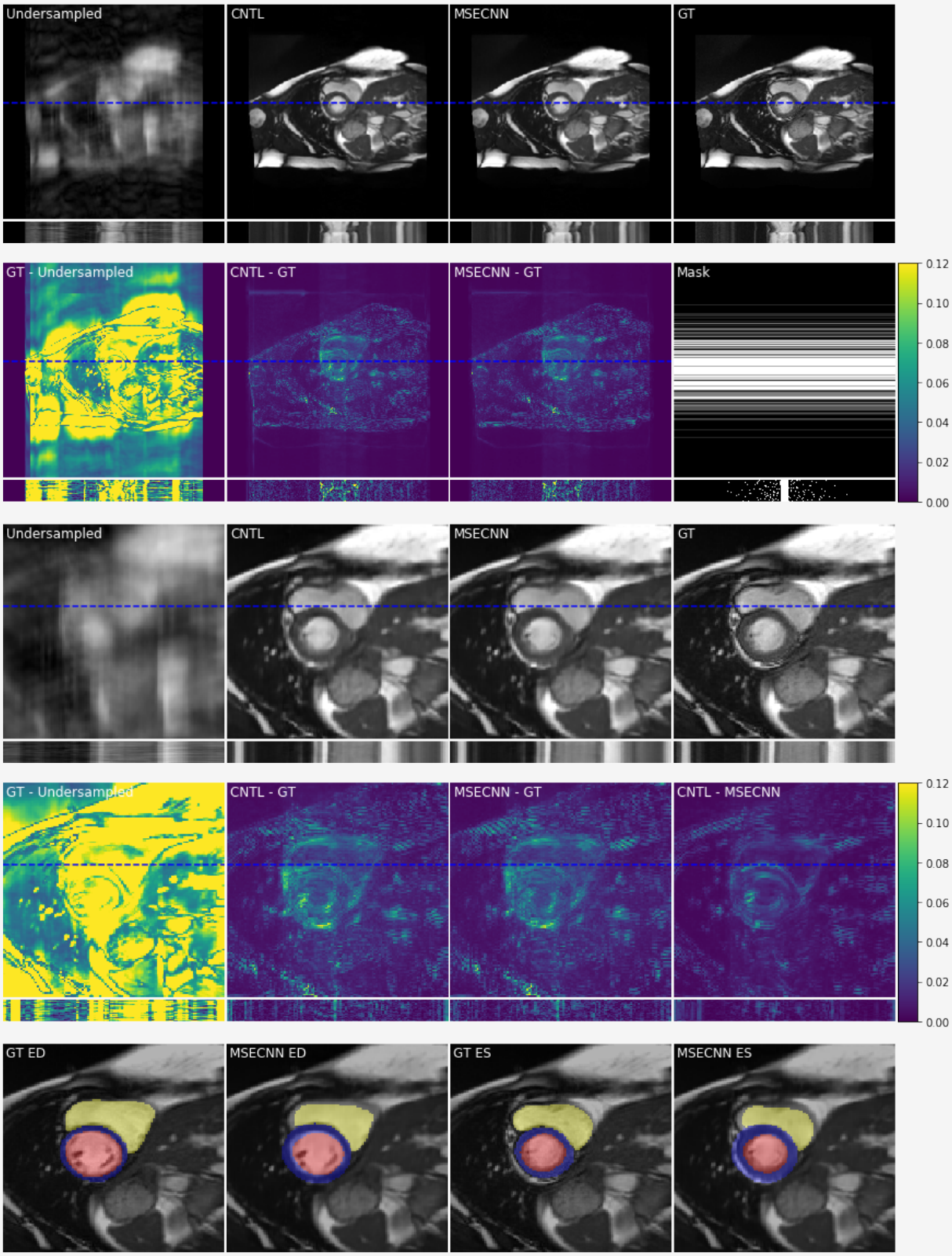

Figure 4.21: Comparison of the reconstruction of a x16 cardiac cine acquisition by the two different reconstruction models. The first and third row are the reconstructed images and enlarged central crop of said images. The second and fourth rows are the difference images with the row above. The last row are the segmentations generated by the MSE-CNN. (PSNR, SSIM) for the reconstructions above: CNTL - (36.5, 0.946), MSE-CNN - (36.9, 0.956). The Dice score of the (ED, ES) segmentations from the MSE-CNN were (0.827, 0.830).

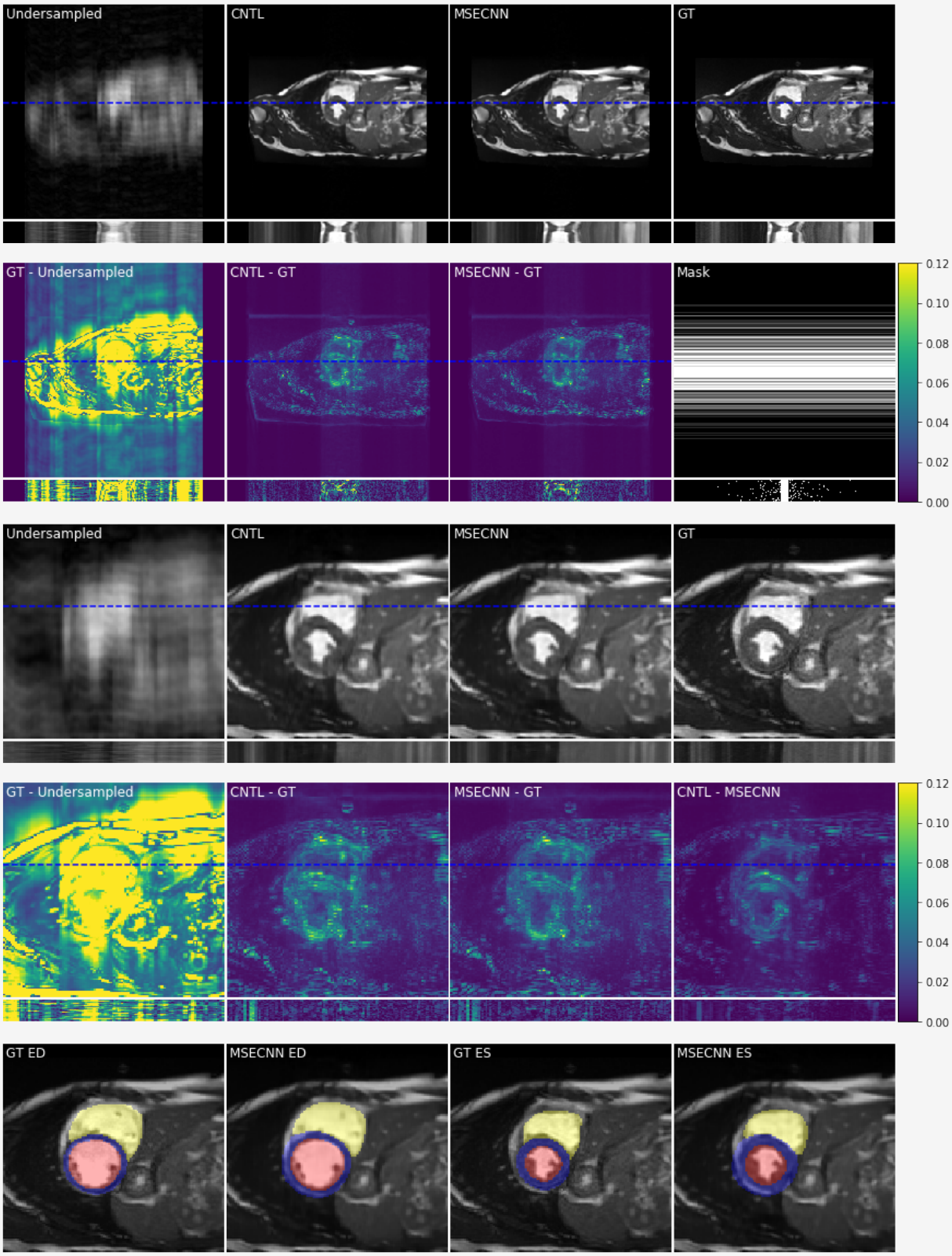

Figure 4.22: Comparison of the reconstruction of a x16 cardiac cine acquisition by the two different reconstruction models. The first and third row are the reconstructed images and enlarged central crop of said images. The second and fourth rows are the difference images with the row above. The last row are the segmentations generated by the MSE-CNN. (PSNR, SSIM) for the reconstructions above: CNTL - (36.1, 0.934), MSE-CNN - (36.5, 0.944). The Dice score of the (ED, ES) segmentations from the MSE-CNN were (0.874, 0.765).

estimate. Furthermore, in order to further highlight the advantage of MSE-CNN, the reconstructions should be investigated at more aggressive acceleration rates in order to the highlight the regularisation benefit of segmentations. Extending this work to training directly on 3D volume cines is also of interest (rather than 2D slices). In particular, in-plane and out-of-plane motion is likely to be the next biggest contributor to PSNR and SSIM gain for MSE-CNN.

4.3 Summary

In chapters 3 and 4, we demonstrated various ways in which deep learning networks can be used to exploit motion estimation. This centered around a decomposition of the optimisation objective in section 3.3 which introduced a motion term. Due to computational requirements of performing motion estimation in conjunction with reconstruction, a number of simplifications were introduced in section 3.5. In this section, it was shown that in a highly accelerated setting, motion exploitation lead to benefits in the reconstruction process in the form of the ME-CNN. In section 3.7, a version of the ME-CNN that more closely followed the decomposition from section 3.3 was implemented by incorporating a recurrent framework and using a single, pretrained crude motion estimator. This was referred to as the ME-CRNN. In Chapter 4, an upgrade of TensorFlow led to the relaxation of some computational restraints that allowed for improvements in the ME-CNN leading to the ME-CNNv2. Using this improved network, a full comparison against state-of-the-art reconstruction algorithms was performed showing the benefit ME-CNNv2 at a range of accelerated rates from x2 to x51.2. In the last section, section 4.2, the use of segmentation data to improve the ME-CNNv2 motion estimate was investigated with the motivation to use the improved motion estimate for enhanced reconstruction. The ME-CNNv2 was modified to share the motion estimator with a segmentation decoder with additional loss functions. Due to computational resource constraints, the motion estimator and segmentation network was shared across all reconstruction blocks. This new network, referred to as the MSE-CNN, performed better reconstructions with segmentation regularisation than without with x16 accelerated acquisitions. This opens a whole new potential area of research for future work in deep learning CMR research.

In conclusion, the benefits of motion estimation in the reconstruction process are clear. However, further work is needed to expand the approaches introduced in chapters 3 and 4 to volume data. Additionally, the framework introduced here can be adapted for motion correction and not just motion exploitation leaving large scope for further exploration.

Chapter 5

Spatial Semantic-Preserving Latent Space Learning for Accelerated DWI Diagnostic Report Generation

In light of recent works exploring automated pathological diagnosis, studies have also shown that medical text reports can be generated with varying levels of efficacy. Brain diffusion-weighted MRI (DWI) has been used for the diagnosis of ischaemia in which brain death can follow in immediate hours. It is therefore of the utmost importance to obtain ischaemic brain diagnosis as soon as possible in a clinical setting. Previous studies have shown that MRI acquisition can be accelerated using variable-density Cartesian undersampling methods. In this study, we propose an accelerated DWI acquisition pipeline for the purpose of generating text reports containing diagnostic information. The model bypasses the traditional image reconstruction step in an effort to streamline the diagnostic pipeline. We demonstrate that we can learn a semantic-preserving latent space for minor as well as extremely undersampled MR images capable of achieving promising results on a diagnostic report generation task. This chapter is based on work completely jointly with equal contribution from Aydan Gasimova and is based on the publication in [159].

5.1 Introduction

Patients that have suffered the symptoms of a stroke have a very short time frame in which to be effectively treated; therefore, it is imperative that radiologists determine the cause of the symptoms in order to provide the appropriate treatment. The majority of strokes are caused by cerebral ischaemia, which can be characterised as reduced blood flow to the brain, causing poor oxygenation that can lead to permanent brain cell death. Both computed tomography (CT) and multi-modal magnetic resonance imaging (MRI) are effective in assessing brain ischaemia, but diffusion-weighted MRI (DWI) is particularly advantageous as it provides highest sensitivity to early ischaemic lesions. In comparison to CT, typical DWI has a much longer acquisition time which additionally makes the scans more susceptible to patient motion and subsequent unwanted imaging artefacts. Furthermore, requiring patients to lay dormant without any motion for long periods of time may lead to discomfort. A well-explored approach for accelerating scan-time is through *undersampling* whereby fewer scanner measurements are taken, violating the Nyquist-Shannon sampling theorem and thus introducing aliasing artefacts into the reconstruction of the image. Several studies are focused on the dealiasing of such images, validating undersampled MRI as an accepted acceleration technique [141, 113, 16, 21, 121, 121, 110, 93].

Assessing the quality of the MR image reconstruction is typically focused on calculating similarity metrics such as peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index between the dealiased reconstruction and the fully-sampled image [26]. This does not, however, guarantee the retention of pathological features necessary for a diagnosis, especially at more aggressive acceleration rates. Therefore, a complimentary way of reviewing extremely accelerated images is through the use of real-time diagnostic tasks such as segmentation and classification [116]. In our study, we explore the automated generation of radiological text reports containing relevant diagnostic and contextual information. The logging of diagnostic reports generated by qualified radiologists is standard hospital protocol. As a result, datasets for studies involving automated text report generation can be acquired directly from hospital archives. In contrast, segmentation and classification tasks require non-standard time-consuming manual annotations. In addition, DWI diagnostic reports typically detail contextual information as well as

the presence/absence of an acute lesion, such as anatomical location and severity of the lesion, and being able to auto-generate them will additionally expedite the process of identifying and documenting acute ischaemia.

To this end we have developed a pipeline that 1) learns an implicit context-preserving manifold of brain DWIs that captures both spatial and pathological information, 2) enforces a latent code for the accelerated DWIs that performs in a similar fashion to the fully-sampled images 3) utilises these accelerated brain DWI image representations to learn to automatically generate reports using a recurrent neural network. To our knowledge, this is the first demonstration of deep latent space learning for the retention of semantic feature information required for accelerated report generation, and the first demonstration of learning to auto-generate reports from brain DWI images.

5.2 Previous work

Latent space learning of accelerated MRI Previous work has shown the use of deep latent space learning for performing tasks such as segmentation and reconstruction in the context accelerated MRI [121, 116]. Accelerated MRI data acquisition is centred around the ability to reconstruct image data in a typically ill-conditioned inverse regression problem. However, certain tasks will only require certain parts of information from the sensor space, called 'k-space'. For example, approximate motion estimation from cardiac cine MRI can be performed with acceleration factors as high as 51.2 [141]. The study in [116] shows that cardiac segmentation can be performed by a single line acquisition in k-space. Inspired by this, we explore the use of deep latent space learning for generating diagnostically-relevant contextual image embeddings. Whilst the study in [116] shows that deep latent space learning provides a manifold that can be robust to different undersampling patterns, they also show that at extreme acceleration rates, deep latent space learning can outperform conventional approaches.

Radiology report generation Learning to automate report generation for radiological images has thus far been heavily influenced by image captioning models formulated as an encoder-

decoder machine translation problem. In image captioning, image representations are extracted from a pre-trained convolutional neural network (the encoder) and passed as inputs alongside captions to a sequence-learning decoder by, for instance, mapping the word and image representations to the same feature space [66, 70]. Such a framework was used by [80] to predict structured medical subject heading (MeSH®) annotations for chest X-ray images.

More recently, learning to attend to spatial visual features has been shown to be effective in image captioning [71] and medical report generation [101, 111, 119, 146]. Using structured reports in a dual-attention framework, Zhang et al.[101] were able to improve features used for classifying histopathology images. The co-attention network of Jing et al. [111] is fed visual as well as semantic features in order to provide high-level semantic information to the text-generation task. Xue et al.[119] break down the task of report generation into subtasks of generating one sentence at a time where each succeeding sentence is conditioned on image features and previous sentences. Yuan et al.[146] also demonstrate the benefit of learning radiology-related features from an initial classification task and go a step further by learning features from multi-view 2-D images (chest X-rays) by introducing a cross-view consistency loss.

The accelerated acquisition of brain DWI has been previously studied in the context of image reconstruction [91, 97, 96, 107]. However, in our study, we explore its use for automated text report generation. We demonstrate how the latent space learned by the accelerated reconstruction network captures both spatial semantic and pathology information required in order to learn to generate reports.

5.3 Method

Our study accelerates DWI acquisition through aggressive variable-density Cartesian undersampling as has been studied in several previous works such as [141, 116]. In our study, we start with attempting a zero-fill reconstruction whereby the lines in k-spaces that are not acquired are filled with zeros. An example of a fully sampled image and a corresponding undersampled,

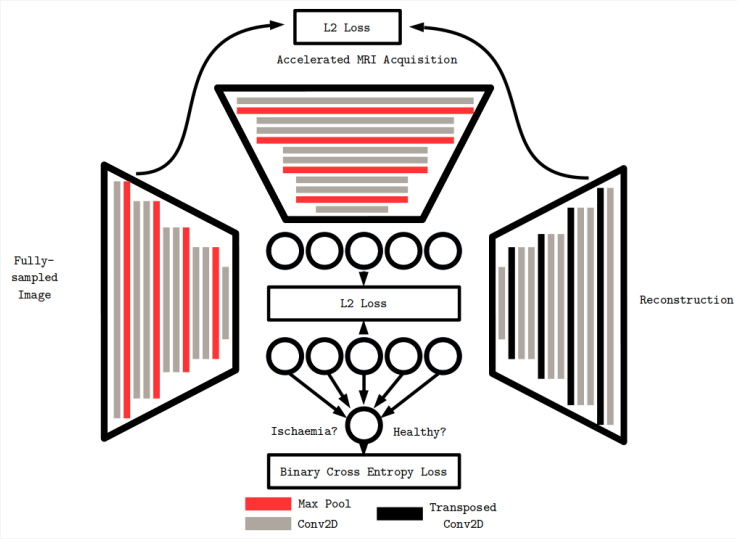

Figure 5.1: An autoencoder is trained to reconstruct the fully-sampled image through an L2 loss. The latent space is conditioned to encode pathological information by performing a classification of ischaemia, trained with a binary cross-entropy loss. The latent space encoding learned at the bottleneck is used as a training target for the encoding branch which only sees the accelerated image.

zero-filled image reconstruction is shown in Figure 5.2. For all acceleration rates, we always sample the two most central lines in k-space whilst the other lines are acquired following a Gaussian distribution centred at the point of highest energy in k-space. During training, undersampling masks are generated on the fly and images are also augmented with additional rotations and translations.

5.3.1 Latent space learning

In our approach, we use an autoencoder network that takes as input the original fully-sampled DWI brain MRI. The purpose of this is to learn a latent space at the bottleneck that contains spatial and contextual information that may be useful for a text report generator. In particular, we manipulate the embedding manifold toward one more suitable for text report generation by introducing an ischaemia-classification loss as a regulariser. This loss can be summarised by equation (5.1) where an Adam optimiser with learning rate 1.0×10^{-5}, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used.

$$L(x, y) = ||D(E(x)) - x||_2^2 - \gamma(y \log C(E(x)) + (1 - y) \log(1 - C(E(x)))), \quad (5.1)$$

where E, D and C are the encoder, decoder and classifier networks (from figure 1) respectively, x is our fully-sampled image, y is a binary classification label for ischaemia and $\gamma = 8000$. We can measure the performance of the latent space learnt as a combination of reconstruction error (in particular of the ischaemia) and of the classification error.

Alongside this, we use a structurally-identical encoding branch to learn a latent space for the accelerated MRI acquisition. We use the approach of performing a zero-filled reconstruction which is passed to a series of convolutional layers. These can be used to identify aliasing artefacts that share information with the unavailable fully-sampled image. In spite of using a heavily aliased image, the generated feature map will consistent of highly relevant image features that are akin to the case whereby the fully-sampled image was used instead. The latent space is trained against the bottleneck of the autoencoder using an L2 loss and another Adam optimizer with the same optimizer parameters. This is summarised in Figure 5.1 and in equation (5.2). Note, for each acceleration rate used in our study, a unique encoder is learned to generate the required latent space. An advantage of deep latent space learning is that we can train the specific encoder associated with different acceleration rates towards the same manifold which avoids the need for retraining of the text report generator model.

$$L(x, x_{\text{acc}}) = ||E(x) - x_{\text{acc}}||_2^2, \quad (5.2)$$

where x_{acc} is our accelerated, aliased image and E_{acc} is our encoding branch for the accelerated images.

5.3.2 Report generation model

We use a report generation model based on [124] where the report word sequence is modelled using the Long Short-Term Memory (LSTM)[10], and conditioned on image embeddings at

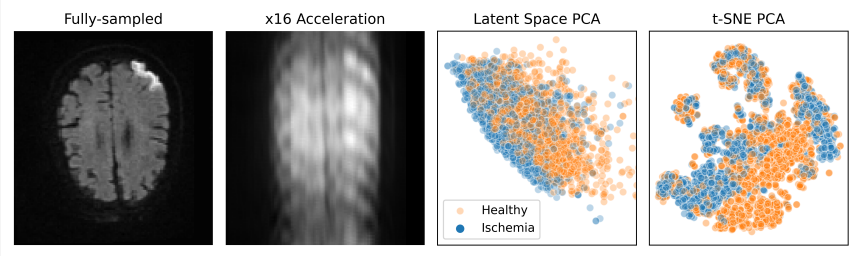

Figure 5.2: Left to right: (1) An example of a brain with ischaemia (2) The corresponding x16 accelerated image is zero-fill reconstructed from k-space using a 2D Fourier Transform. Note that this image is infected with heavy aliasing artefacts. (3) A projection of the first two principle components in a PCA analysis of the latent space. Some clustering can be seen (4) a t-SNE projection of the latent space showing clear clustering.

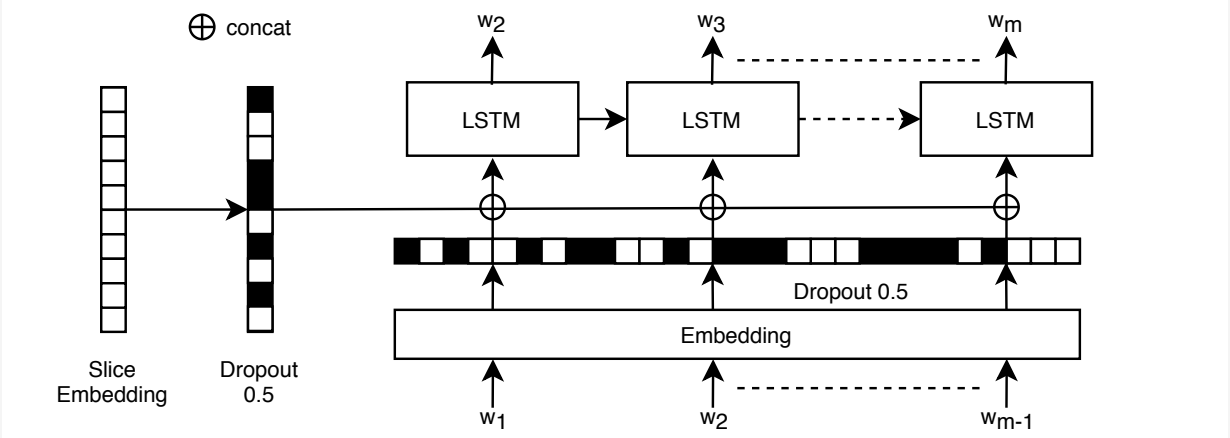

Figure 5.3: Clinical report generation model from accelerated image latent space embeddings.

each time step through concatenation at the input to the LSTM. At each time step, the input, output and forget gates control how much of the previous time steps is propagated through to the output. For an input embedding sequence $\{x_1, \ldots, x_n\}$ where $x_i \in \mathbb{R}^D$, the internal hidden state $h_t \in \mathbb{R}^h$ and memory state $m_t \in \mathbb{R}^m$ are updated as follows:

$$\begin{aligned} h_t &= f_t \odot h_{t-1} + i_t \odot \tanh(W^{(hx)} x_t + W^{(hm)} m_{t-1}) \\ m_t &= o_t \odot \tanh(h_t) \end{aligned} \tag{5.3}$$

where $x_t \in \mathbb{R}^D$ is the concatenation of the latent space image embedding and word embedding at time step t, $W^{(hx)}$ and $W^{(hm)}$ are the trainable weight parameters, and i_t, o_t and f_t are the input, output and forget gates respectively. The model architecture is illustrated in Figure 5.3. We additionally add Dropout layers after image and word embeddings to force the model to condition on both thus regularising training.

5.4 Experiments

The Data The dataset consists of 1226 3D DWI scans and corresponding radiological reports of acute stroke patients. All the images and reports were fully anonymised and ethical approval was granted by Imperial College Joint Regulatory Office. The scans were pre-processed according to the steps outlined in [85]: images were resampled into uniform pixel size of 1.6×1.6mm, and pixel intensities were normalised to zero mean and unit variance. The number of slices per image varies between 7 and 52, and the slice dimensions are 128×128.

Each report contains between 1 and 2 sentences summarising the presence or absence of the pathology, a visual description, and its location within the brain. In addition, each exam is assigned a diagnostic label as part of hospital protocol: 54% were diagnosed 'no acute infarct', 46% were diagnosed 'acute infarct'. The remaining, which made up a total of <1% and included diagnoses such as 'unknown', 'haematoma', 'tumour', were removed for the purpose of training. Processing was done on the reports to remove words outside the 99th percentile, exams with

empty reports were removed, leaving a total of 1104 exams, total vocab length 1021, mean words per exam 10.8, std. 6.3.

In order to simplify the problem, we created a 2D dataset of acute and non-acute (normal) slices from these images. For the acute set, we used the brain ischaemia segmentation network developed by Chen et al.[85] to segment the images labelled with acute ischaemia, thresholded at 0.8, and selected slices where the total area of ischaemia was >10 pixels. For the normal set, we sampled slices from the non-acute labelled images according to the same axial plane distribution as the acute set.

Experimental settings Reports were padded with 'start' and 'end' tokens to length 19 (mean + 1std. + 'start' + 'end'). The word embedding layer maps one-hot encoded word embeddings into a learnable 256 dimensional space. The LSTM hidden state is also set to dim 256, and the LSTM units are unrolled up to 19 time steps. We train the model on non-accelerated latent embeddings and their associated reports by minimising the categorical cross-entropy loss over the generated words. All models are trained with batch size 128, using Adam optimisation [61], learning rate=0.0001 for 300 epochs. The language model in total had 1.45M parameters and the separately trained model that generated the latent embeddings had 126M parameters.

Results Inference was performed by first sampling from the LSTM using a 'start' token concatenated with the accelerated embeddings, and consequently appending the output word embedding to the input and sampling until an 'end' token was reached. The quality of the generated reports was evaluated by measuring BLEU [22] and ROUGE [23] scores averaged over all the reports, which are a form of n-gram precision commonly used for evaluating image captioning as they maintain high correlation with human judgement. We observe that the both the BLEU and ROUGE scores decrease with increasingly accelerated images, as expected. We note that there is a significant reduction in performance between the x4 and x8 accelerated images possibly due to some contextual information not being captured by the latent space.

We also assess the sampled reports qualitatively in Figure 5.4. We observe no major grammatical errors for all accelerations, an no major content errors for lower accelerations with x2 and x4 correctly identifying the presence/absence of ischaemia as well as the location. Note: the

Table 5.1: BLEU1,2,3,4-gram and ROUGE1 f1, precision (P) and recall (R) metric comparisons on increasingly accelerated image embeddings.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1 F1	ROUGE-1 P	ROUGE-1 R
Acc.×1	38.12	27.26	20.28	15.59	47.10	52.89	44.96
Acc.×2	34.07	23.31	15.55	11.57	44.00	51.86	40.68
Acc.×4	31.36	19.42	12.29	8.31	41.17	48.09	38.80
Acc.×8	21.32	10.37	5.06	2.55	29.53	32.92	29.52
Acc.×64	21.58	11.11	4.97	2.35	30.39	35.10	29.07

last example shows a text report that was ischaemic but was classified as healthy. This is likely to have confused the latent code for this example resulting in poor text report generations.

5.5 Extension to 3D volume data

In our study above, we used 2D slices of the k-space acquisition in the model that generate the latent code. This is a sub-optimal methodology for re-use with a different dataset due requirement of a complex segmentation model to identify ischaemic slices for balacing the training dataset. It should be noted that there are far more ischaemic slices that normal/non-ischaemic. Instead, we propose using the 3D data directly without a separate slice extraction pipeline, making the whole process more streamlined. We expand on [159] to generate radiological texts from 3D pathological volumes without an intermediate data cleaning stage. We 1) Build a latent encoder balanced by the auxiliary tasks of classification and image reconstruction without accelerated acquisition 2) Tune our encoder through a set of experiments with a validation set on a report generation task 3) Train our final, tuned model with accelerated acquisitions for DWI report generation without an intermediate reconstruction phase. The hypothesis of this study is the same as in the previous section - that using a combination of these auxiliary tasks will help to better retain contextual information in the resulting latent encoding which will lead to a subsequent increased performance for accelerated radiological text report generation, just as in the section above.

Hyperparameter search One goal of our study was to ascertain the balance between our

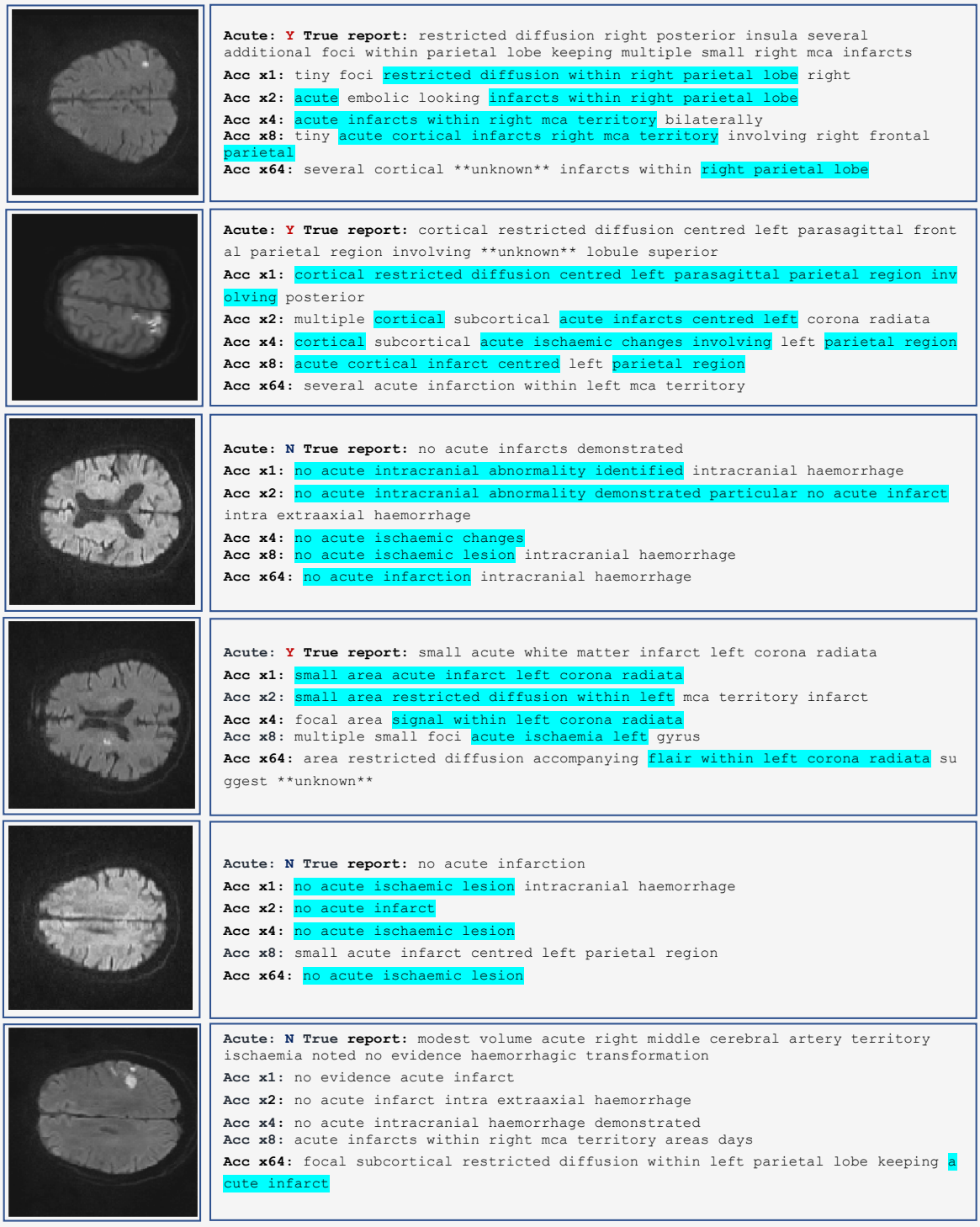

Figure 5.4: Sample brain slices and associated reports generated from non-accelerated and increasingly accelerated image embeddings. Correctly identified pathology (acute/non-acute) and spatial contexts are highlighted in blue.

Table 5.2: Results of hyperparameter search

Model	Acc.	Precision	Recall	B-1	B-2	B-3	B-4
Classification Only	0.79	0.85	0.67	**21.10**	8.73	4.14	0.47
$\gamma = 1e8$	0.54	0.50	0.53	12.28	3.02	2.10	0.00
$\gamma = 1e9$	0.62	0.58	0.60	18.49	9.02	1.85	0.70
$\gamma = 1e10$	0.78	0.82	0.65	20.83	**11.82**	**8.62**	**7.59**
$\gamma = 1e11$	0.77	0.76	0.73	18.68	9.88	2.65	1.32

auxiliary tasks for the latent space learning from 3D volumes and thus optimise the parameter γ. We assess the quality of the latent space by training and then sampling from the report generation model, and evaluating the predicted reports against the true reports using the BLEU metric, a modified n-gram precision metric, averaged across samples. The results are shown in Table 5.2. After evaluating on the validation set, we found that the classification only model performed best on BLEU-1, however, when $\gamma = 1e10$, the model performs better on higher n-gram BLEU metrics. Higher BLEU metrics on longer n-grams indicates that a more contextual report is learned (i.e. greater overlap of 2, 3, and 4 sequential words). This is consisted with our hypothesis that the auxiliary task of reconstruction improves the semantic-preserving ability of the latent space.

Accelerated DWI report generation With the optimal hyperparameters chosen for the auxiliary learning task, the 'accelerated' encoder was trained to produce the same embeddings of 'fully-sampled' encoder via an L2 loss. The result was that the semantic embeddings were produced from extremely accelerated acquisitions of pathological brain volumes. We found that even highly accelerated acquisitions were able to be encoded to representations very close to that of fully-sampled acquisitions. These embeddings were then used to produce the associated accelerated radiological text report. The BLEU scores evaluated on the test dataset for each acceleration rate is shown in Figure 5.6. As expected, higher acceleration rates lead to worse BLEU scores but it is important to note that even at x8 acceleration, the reports are still of good quality as shown in Table 5.3.

Table 5.3: Sample ground truth and generated reports from fully sampled and undersampled 3D brain DWI. Correctly identified concepts are highlighted.

True:	'no', 'acute', 'ischaemic', 'lesion', 'intracranial', 'haemorrhage'
No Acc.:	'no', 'acute', 'infarct', 'intra', 'extraaxial', 'haemorrhage', 'demonstrated'
Acc. x8:	'per', 'mri', 'study', 'performed', 'earlier', 'today', 'no', 'acute', 'intracranial', 'abnormality', 'evident'
True:	'multiple', 'small', 'acute', 'infarcts', 'scattered', 'throughout', 'left', 'superior', 'temporal', 'inferior', 'frontal', 'superior', 'parietal', 'lobe'
No Acc.:	'acute', 'cortical', 'left', 'mca', 'territory', 'infarct', 'within', 'left', 'parietal', 'lobe'
Acc. x8:	'appear', 'small', 'acute', 'left', 'left', 'superior'
True:	'restricted', 'diffusion', 'involving', 'left', 'posterior', 'temporal', 'lobe', 'external', 'capsule', 'posteriorly', 'extending', 'left', 'parietal', 'lobe', 'appearances', 'keeping', 'acute', 'left', 'mca', 'infarct'
No Acc.:	'several', 'small', 'foci', 'restricted', 'diffusion', 'within', 'left', 'parietal', 'lobe', 'keeping', 'acute', 'right', 'mca', 'territory'
Acc. x8:	'minor', 'microangiopathic', 'ischaemic', 'changes', 'involving', 'left', 'occipital', 'lobe', 'extending', 'posterior', 'internal', 'capsule'
True:	'no', 'acute', 'infarction', 'intracranial', 'haemorrhage'
No Acc.:	'no', 'acute', 'infarct', 'haemorrhage', 'demonstrated'
Acc. x8:	'no', 'acute', 'infarct', 'evidence', 'recent', 'haemorrhage', 'demonstrated'
True:	'acute', 'infarcts', 'seen', 'left', 'frontal', 'corona'
No Acc.:	'acute', 'infarct', 'left', 'corona', 'radiata', 'involving', 'june', 'posterior', 'limb', 'left', 'internal', 'capsule'
Acc. x8:	'acute', 'infarct', 'left', 'corona', 'radiata'
True:	'acute', 'infarction', 'right', 'mca', 'territory', 'involving', 'caudate', 'nucleus', 'anterior', 'limb', 'internal', 'capsule', 'entire', 'lentiform', 'nucleus'
No Acc.:	'complete', 'right', 'aca', 'mca', 'territory', 'infarcts'
Acc. x8:	'note', 'made', 'extensive', 'right', 'mca', 'territory', 'subacute', 'infarct', 'involving', 'right', 'corpus', 'striatum', 'corona', 'radiata', 'external', 'capsule', 'insular', 'right', 'frontoparietal', 'cortices', 'confluent', 'large', 'infarct'

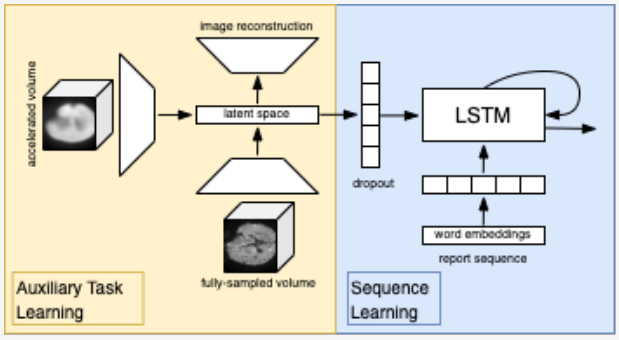

Figure 5.5: Model architecture

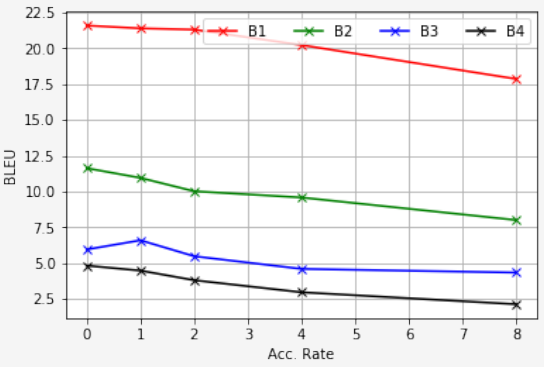

Figure 5.6: Average BLEU-n scores of accelerated brain volumes.

5.6 Cleaning the text reports dataset

As a separate investigation, the text reports in the training set were more dramatically cleaned. The vocabulary was reduced to just 113 words (from 1021) and each sentence was manually inspected and reworded. Examples of this are shown in Table 5.4. We hope to gain better and more consistent text report outputs which will allows us to better ascertain how the quality of the text reports degrade with acceleration rate, particularly in terms of 3-gram and 4-gram BLEU.

Volume data In this setup, we processed the data slightly different. A volume in the dataset was chosen and removed so that it could be used a reference volume. We then used an affine transformation followed by a B-spline free-form deformation to register the volume to the reference [17]. The data volume, x, is then normalised between 0 and 1. The normalisation takes place by finding the 81.25 percentile, x_l, and the 93.75 percentile, x_u, of all voxels within a given volume. These values encapsulate the lower and upper bound of meaningful values in the data. The maximum value at which clipping occurs is $\max x_u + 4 * (x_u - x_l), \max x$. The minimum value at which clipping occurs is the $x_l - 0.5(x_u - x_l)$.

Results Using 3-fold cross validation, the hyperparameter $\gamma = 10^{\phi}$ was found to be optimal in terms of all n-gram BLEU scores for $\phi = 4.0$, outperforming a classification only model as seen in Figure 5.7. We then evaluate the model on the test set against a range of different

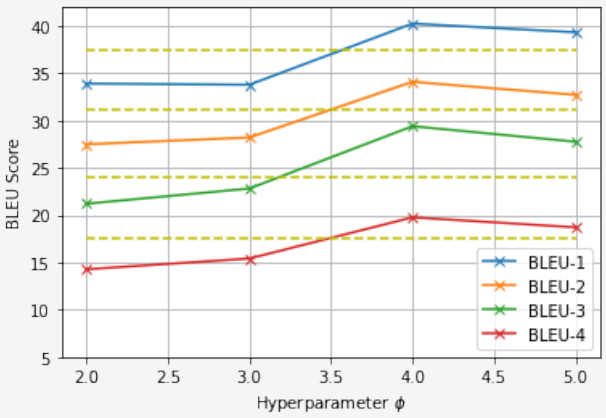

Figure 5.7: BLEU scores with varying hyperparameter that balances the reconstruction and classification capabilities of the latent code. The dotted line shows the BLEU scores of a classification only model, with the highest dotted line being BLEU-1.

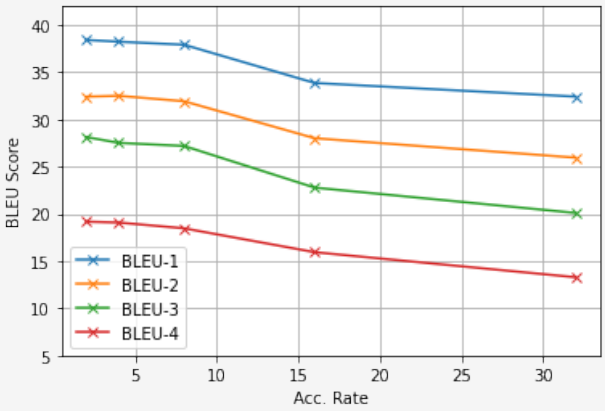

Figure 5.8: Performance of report generation model with increasing acceleration rate. The report generation is robust against acceleration rate until as aggressive as $\times 8$ acceleration.

acceleration rates from $\times 2$ to $\times 32$ as shown in Figure 5.8. Some examples of generate reports are shown in Table 5.4. We confirm that results found in the previous sections but with better context in the generated text reports as well as a more stable training process.

5.7 Conclusion and future work

We demonstrate how a latent space capturing pathalogical and spatial information can be learned from accelerated brain DWI images and subsequently used to train a diagnostic report generation network with promising results.

We present a streamlined pipeline that directly transforms an accelerated DWI acquisition into a semantically-rich embedding space, from which radiological text reports can be learned. Another aim of this preliminary study was to ascertain the use of balanced reconstruction and classification auxiliary tasks for the generation of image embeddings in the context of accelerated radiological report generation. Overall, we demonstrate how a balanced latent space capturing pathological and spatial information can be learned from accelerated brain DWI images and

Table 5.4: Sample ground truth and generated reports from fully sampled and undersampled 3D brain DWI using the cleaned text reports. Correctly identified concepts are highlighted.

Reference:	there is no acute infarct and there is no intraaxial haemorrhage and there is no extraaxial haemorrhage
Acc. Rate 2:	there is no acute infarct and there is no haemorrhage
Acc. Rate 4:	there is no acute ischaemic lesion and there is no haemorrhage
Acc. Rate 8:	there is no acute infarct and there is no intracranial haemorrhage
Acc. Rate 16:	there is no acute ischaemic lesion
Acc. Rate 32:	there is no acute infarct and there is no intracranial haemorrhage
Reference:	there is a acute infarct in the right corona radiata and in the posterior thalamus capsule region
Acc. Rate 2:	there is a foci of restricted diffusion within the right corona radiata suggesting there is a acute infarct
Acc. Rate 4:	there is a foci of restricted diffusion within the right posterior lateral thalamus and there is a signal abnormality suggesting there is a acute infarct
Acc. Rate 8:	there is a signal abnormality suggesting there is a acute infarct and there is a foci of restricted diffusion within the right corona radiata suggesting there is a acute infarct
Acc. Rate 16:	there is a signal abnormality suggesting there is a acute infarct and there is a foci of restricted diffusion within the right posterior lateral thalamus
Acc. Rate 32:	there is a foci of restricted diffusion within the right posterior lateral thalamus and there is a signal abnormality suggesting there is a acute infarct
Reference:	there is a subacute infarct in the left middle cerebral artery territory
Acc. Rate 2:	there are multiple acute infarct in the left middle cerebral artery territory
Acc. Rate 4:	there is a foci of restricted diffusion in the right frontal parietal region and there is signal abnormality
Acc. Rate 8:	there are multiple foci of signal abnormality within the deep white matter
Acc. Rate 16:	there is a large haemorrhage infarct in the left middle cerebral artery territory
Acc. Rate 32:	there is a acute infarct in the left insular lobe and in the left middle cerebral artery territory and in the right middle cerebral artery territory
Reference:	there is a large restricted diffusion and there is a signal abnormality suggesting there is a acute infarct in the posterior left middle cerebral artery territory
Acc. Rate 2:	there is a acute infarct in the right middle cerebral artery territory and in the insular right the right insular lobe and in the left insular lobe
Acc. Rate 4:	there is restricted diffusion in the left inferior parietal lobe suggesting there is a acute infarct in the left middle cerebral artery territory and in the left frontal operculum and in the left parietal lobe
Acc. Rate 8:	there are multiple foci of cortical acute infarct in the left parietal lobe and in the left frontal lobe
Acc. Rate 16:	there are multiple cortical and subcortical infarct in the left precentral gyrus and in the postcentral gyrus and there is a infarct in the left occipital parietal region
Acc. Rate 32:	there is a infarct in the left middle cerebral artery territory
Reference:	there are multiple foci of restricted diffusion and there is signal abnormality in the medial right occipital lobe and there are multiple foci within the thalamus bilaterally suggesting there is acute ischaemia
Acc. Rate 2:	there is a subacute right posterior cerebral artery territory infarct
Acc. Rate 4:	there is a acute right posterior cerebral artery territory infarct in the right occipital lobe and in the right splenium of the corpus callosum and in the right lateral thalamus and in the medial right temporal lobe
Acc. Rate 8:	there is a subacute right posterior cerebral artery territory infarct
Acc. Rate 16:	there are multiple foci of acute ischaemia in the right posterior medial temporal lobe and in the posterior circulation and in the right posterior medial temporal lobe and in the cerebral hemisphere and in the right thalamus and in the posterior limb of the internal capsule
Acc. Rate 32:	there is a acute infarct in the right pons and there is restricted diffusion within the right corona radiata suggesting there is a small acute infarct

subsequently used to train a diagnostic report generation network with promising results.

Future progress from this preliminary study includes investigations into different acceleration schemes (e.g. Parallel Imaging) and more sophisticated language models. We also wish to explore radial undersampling trajectories for DWI brain imaging which are expected to provide improved diagnostic embeddings.

Broader Impact We envision this work being used in a streamlined diagnostic pipeline. At the moment, during accelerated acquisition, we can expect to see a reconstructed image and perhaps its associated pathological segmentation. However, with further development, we wish to see this work being to generate pathological text reports directly from the accelerated acquisition, particularly in cases where immediate diagnostic information is needed such as with brain ischaemia where brain death can be imminent if not treated immediately.

Acknowledgments

We would like to thank Dr Paul Bentley, Clinical Director of the Imperial College Network Of Excellence, for providing us with access to the DWI dataset.

Chapter 6

ME-DDPM: Motion Exploiting Denoising Diffusion Probablistic Models

Diffusion Models are a class of generative models that have shown to produce high-fidelity image reconstructions from accelerated MRI acquisitions compared to conventional reconstruction algorithms. In the case of dynamic data — for example, cardiac cine — aggressive acceleration rates for image reconstruction have been realised with the advent of a recent method called the ME-CNN that uses motion estimation for temporal data exploitation. In this work, we show that diffusion models provide a natural way to use temporal data exploitation to better guide diffusion processes towards the true intermediate distribution resulting in higher quality reconstructions. Inspired by the ME-CNN, we introduce a new class of guided diffusion models for cine MR reconstruction. Our proposed model, ME-DDPM, is compared against suitable baselines that suggests that ME-DDPMs can be used to help aggressively accelerate cine acquisitions with competitive fidelity.

6.1 Introduction

A new class of generative models called diffusion models have recently made high fidelity image generation possible without the cumbersome training and mode collapse typically experienced with GANs [183, 162, 86, 84, 108]. The flavour of diffusion model we investigate in our study is called the denoising diffusion probablistic model (DDPM) which is also a variance preserving model [1] [170]. However, the methods we introduce can be generalised to any flavour of diffusion model.

Diffusion models can be interpreted as reversing a stochastic process with a predefined drift and volatility/variance. [161, 170] show that for a diffusion model with a certain variance schedule and a large number of diffusion steps, the forward diffusion process transforms our data distribution to approximately $N(0, I)$. A key feature of the DDPM is that we can easily access every latent distribution under certain choices of the variance schedule of the diffusion process. This allows for an efficient training process. The trained DDPM model can then be used to approximate the data distribution so that we may freely sample from the DDPM to generate new images. We direct the reader to section 2.4 for more information on diffusion models.

Concurrently, there has been work relating the motion exploitation of dynamic sequences to aid in the reconstruction process ([141], chapters 3 and 4). Inspired by our previous work, we propose to use similar motion exploitation for reconstructing dynamic sequences but in the context of diffusion models. DDPMs offer advantages in high fidelity reconstructions but possibly in aleatoric uncertainty estimation and handling noisy data (see section 7.5 for more information on the use case of DDPMs). Combining motion exploitation with DDPMs allows us to obtain even higher fidelity reconstructions under aggressive acceleration rates.

[1] Equations 5.13 and 5.51 in [136] give $d\text{Cov}(x)/dt = 0$ as opposed to $\propto t^2$ for variance exploding models.

6.1.1 Diffusion Models

Recent work has focused on modelling data generation as a stochastic process [143, 169, 170, 193, 161]. Typically, this involves transforming every data point in our target distribution $q(\mathbf{x})$ to $N(\mathbf{0}, \mathbf{I})$. The stochastic differential equation that models this evolution is shown in equation (6.1):

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, k)dk + g(k)d\mathbf{w}, \tag{6.1}$$

where x is data point being transformed, $\mathbf{w}$ is a Wiener process, k is the process index, $\mathbf{f}(\cdot)$ is the drift term and $g(\cdot)$ is the volatility/variance term . This process is known as a diffusion process as it has known solutions for particular choices of the functionals $\mathbf{f}(\mathbf{x}, k)$ and $g(k)$. Performing this transformation in reverse would instead gradually map $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$ to $q(\mathbf{x})$. The concept is similar to conventional GANs [60] but the process by which this occurs is closer to that of normalising flows whilst also bearing similarities to denoising autoencoders. This reverse transformation is also a diffusion process and can be written as equation (6.2) [1, 170]:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, k) - g(k)^2 \nabla_x \log q_k(\mathbf{x})]d\tilde{k} + g(k)d\tilde{\mathbf{w}}, \tag{6.2}$$

where $\tilde{\mathbf{w}}$ and $\tilde{k}$ are in the reverse direction to the process index k and $q_k(\mathbf{x})$ is the distribution of our data after it has been diffused with equation (6.1) until step k. The gradient of the log data density from equation (6.2), $\nabla_x \log q_k(\mathbf{x})$ — also known as the score, $\mathbf{s}(\mathbf{x}, k)$ — is hard to compute. Instead, we use a neural network (NN) to learn to calculate this — the NN learns the noise present in the current diffusion image, $\epsilon_\theta(\mathbf{x}_k, k)$ which is a proxy for learning the score as $\epsilon_\theta(\mathbf{x}_k, k) = \mathbf{s}(\mathbf{x}, k)/\sqrt{1 - \tilde{\alpha_k}}$ [196]. θ are the parameters of the NN.

In this study, the drift and volatility terms in equation (6.1) are set to match the set-up in study by [161]. This results in a reverse diffusion step as shown in equation (6.3).

$$\mathbf{x_{k-1}} = \frac{1}{\sqrt{\alpha_k}}\Big(\mathbf{x_k} - \frac{1-\alpha_k}{\sqrt{1-\tilde{\alpha}_k}}\epsilon_\theta(\mathbf{x}_k, k)\Big) + \sigma_k \mathbf{z}, \tag{6.3}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\sigma_k = \sqrt{\beta_k}$, $\tilde{\alpha}_k = \prod_k (1-\beta_k)$, $\alpha = 1 - \beta_k$ $\beta_k = \beta_0 + \frac{(\beta_K - \beta_0)}{K} k$ where K is the chosen number of diffusion steps in the model and $\beta_0, \beta_K \ll 1$ controls the prescribed variance schedule.

We advise the reader to see section 2.4 and [161] for more details on training our chosen flavour of diffusion model for this study.

6.1.2 MRI Reconstruction

The problem of MRI reconstruction, equation (6.4), can be written as an unrolled optimisation that consists of successively applied denoising and data consistency steps, the latter of which has a closed form solution in the form of equation (6.5) for single coil acquisitions [49].

$$\frac{\lambda}{2}\sum_t ||Ex_t - y_t||_2^2 + \mathcal{R}(x_0, ..., x_T), \tag{6.4}$$

where E is the encoding matrix that usually consists of an undersampling mask, D, and a Fourier transform $\mathcal{F}$, y_t is the acquired k-space data at cine time frame t, T is the total number of temporal frames in the acquisition and $\mathcal{R}$ is the regularisation imposed on the reconstruction.

$$x^i = (\lambda \mathcal{F}^T D^T D \mathcal{F} + I)^{-1}(\lambda \mathcal{F}^T D^T y + \mathrm{CNN}_k(x^{i-1})), \tag{6.5}$$

where i is the index of the iterative reconstruction process and λ controls the trade-off between the acquired data and the CNN regulariser (in the noiseless case, $\lambda \to \infty$). More information regarding this solution can be found in section 3.3 and in [49, 93].

Similarly, it can be shown that dynamic MRI reconstruction may be written as an unrolled op-

timisation that consists of a 'DCMAC'[2] or motion augmentation term which introduces motion in the reconstruction optimisation process (see chapters 3 and 4 for more information; [141]). One possibility of introducing motion into the optimisation is via the temporal consistency term in equation (6.6).

$$\frac{\lambda}{2}\sum_t ||D\mathcal{F}x_t - y_t||_2^2 + \frac{\rho}{2}\sum_t ||D\mathcal{F}M_t x_t - y_{t+1}||_2^2 + \mathcal{R}(M, x_0, ..., x_T), \tag{6.6}$$

where M_t is the motion operation that transforms an image from time frame i to $i+1$ and ρ controls the trade-off between the DCMAC term and the regularisation.

A decomposition of this optimisation results in two closed forms expressions — a data consistency (DC) term [49] and a DCMAC term (see below and [141]) — and two proximal mapping terms that correspond to the image denoiser and a motion estimator (off loaded to CNNs). We advise the reader to see chapter 3 for more information on this decomposition.

The DCMAC term is shown in the following equation:

$$x_i^k = (\lambda\mathcal{F}^T D^T D\mathcal{F} + \rho I)^{-1}(\lambda\mathcal{F}^T D^T y_{t+1} + \rho M_t^k u_t^{k-1}), \tag{6.7}$$

where $u^{k+1} = \text{CNN}_\theta(u^k, x^k)$. In the case of a perfect, noiseless MRI scanner ($\lambda \to \infty$) and perfect motion estimation, the DCMAC step involves taking the output from the previous cascade/iteration, the denoised frame, and warping it to the next frame where data consistency is subsequently applied.

In our study, we suggest that the conventional CNN can be replaced by that of a DDPM with two main motivations behind this: (1) Increased fidelity whilst having the ability to sample from the learned data distribution (2) Direct manipulation of the latent representations of the DDPM to better leverage our acquired (accelerated) data. In particular, we take inspiration from the data consistency term and equation (6.7) — the DCMAC term — to allows us to direct our latent representations towards one that contains information from the exact distribution

[2]Data-Consistent Motion-Augmented Cine

for the acquired data whilst also exploiting motion information to leverage the entire temporal direction of the acquired data.

6.1.3 Langevin Diffusion for refinement of latent estimate

Introducing stochastic differential equation (6.8), it can be shown that the evolution of the sampling distribution x converges to the stationary distribution $q_k(x)$ using the Fokker-Planck equation [193, 170]. As a result, this Langevin-like diffusion process samples from $q_k(x_k)$ and can be used to refine estimates of our sample x_k with a carefully chosen step size, δ. This method has been referred to as the 'corrector' method in previous work [170, 180].

$$dx_k = -\nabla \log q_k(x_k) d\tau + \epsilon dW, \tag{6.8}$$

where $x_k \sim q_k(x_k)$, W is a Wiener process and τ is the temporal index of said process. Using the Euler-Maruyama discretisation, we can approximate Langevin diffusion with our trained score function as in equation (6.9).

$$x_k^{\tau+1} = x_k^{\tau} - \delta \frac{\epsilon_\theta(x_k^{\tau}, k)}{\sqrt{1-\tilde{\alpha_k}}} + \sqrt{2\delta} z, \tag{6.9}$$

where τ denotes the iteration of the corrector method, δ is the step size used in the discretisation, and $z \sim \mathrm{N}(0, \mathrm{I})$. This is presented in algorithm 2.

6.2 Related Work

Motion-based dynamic MRI reconstruction has been studied with conventional optimisation techniques, notably kt-FOCUSS introduces a way for motion to be used in aiding the reconstruction process given fully-sampled references [42]. Also requiring fully-sampled references is a recently proposed deep learning end-to-end motion guided reconstruction network called MODRN [190]. This resembles kt-FOCUSS with motion estimation and motion compensation

(ME/MC; see section 4.1.3 for more information) but using CNNs for regularisation (U-net with recurrent bottlenecks). These techniques require fully-sampled, high quality reference frames which may not always be attainable such as with certain patients or in the case of fetal cardiac data.

More recently, there has been extensive work that was first to introduce motion estimation into the problem of deep learning reconstruction (see chapters 3 and 4; [141]). This takes the form of the end-to-end ME-CNN (motion exploiting CNN) whereby motion is used to generate data consistent motion-augmented cines (DCMACs) that aid in exploiting data in the temporal direction. The decomposition in section 6.1.2 is based on this work.

With diffusion models, previous work in [194, 180] show that data consistency can be used in the reverse diffusion process to guide the latent representation towards performing MRI reconstruction from undersampled acquisitions. Interestingly, [180] show that a diffusion model trained to generate magnitude images can be used for complex valued acquisitions without having seen complex data at training time. At the time of writing, there is no work that incorporates motion into the DDPM framework.

6.3 Method

In order to integrate the DC and DCMAC term into the ME-DDPM, we acknowledge that the DCMAC term is usually followed by a denoising term that is usually off-loaded to a vanilla CNN at training time. In our proposal, we propose no changes to the current training of DDPMs on cine MRI data. Given the diffusion variable x_k^t at diffusion step k (sometimes referred to as latent representation), we apply the DCMAC term using the correct scaling of the data for the given variance at diffusion step k. Instead of applying a conventional CNN denoiser after application of the DCMAC term, we instead use Langevin diffusion to effectively denoise the latent representation x_k^t. The DCMAC term introduces new information from time frame $t-1$ into time frame t via motion exploitation. As a consequence, an interpretation of Langevin diffusion after the DCMAC step is to ensure that the new information introduced matches the

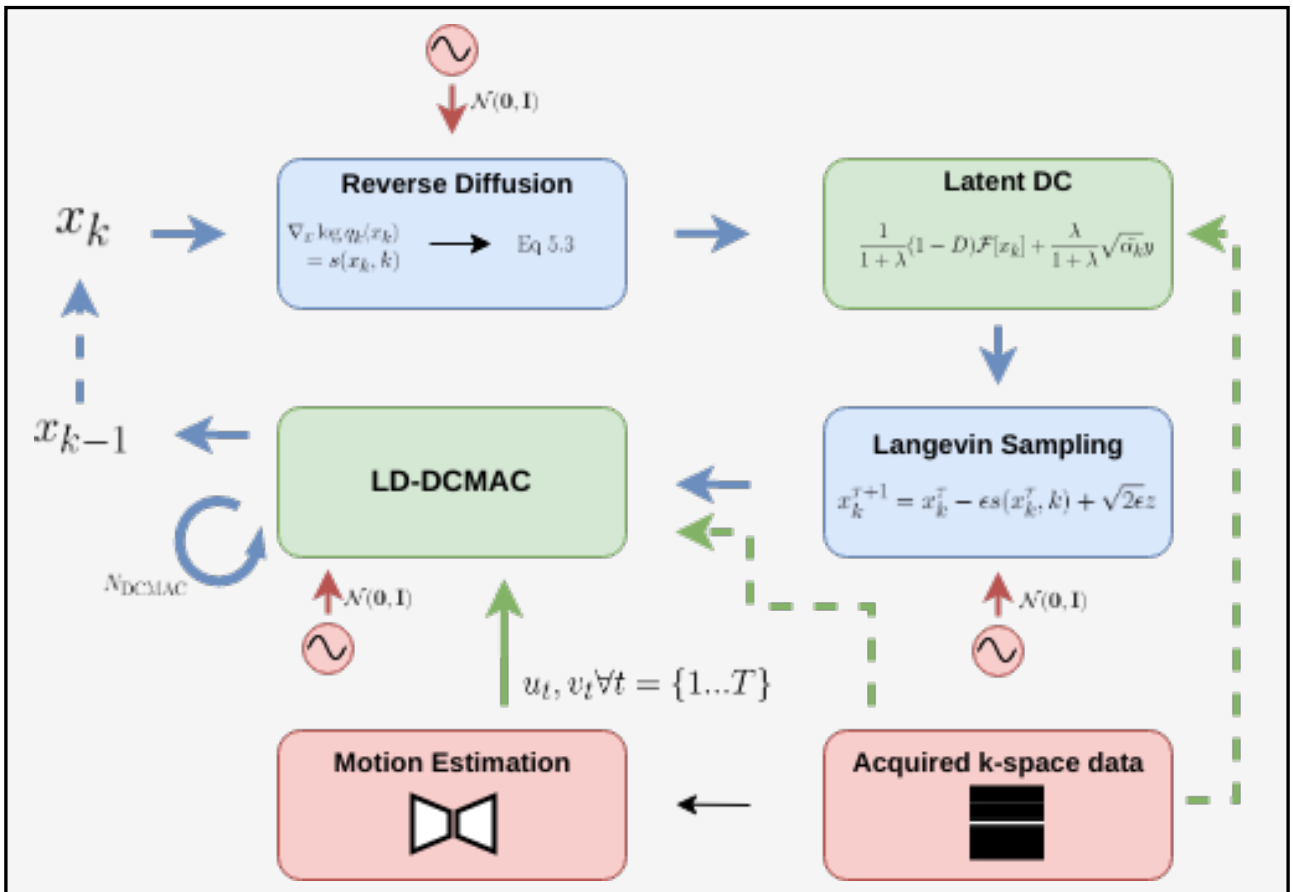

Figure 6.1: Our proposed ME-DDPM. The process starts with x_K drawn from a standard normal distribution. The latent variable, x_k, is reverse-diffused followed by data consistency and our LD-DCMAC block. This ultimately generates an estimate for a new latent variable for the next reverse diffusion step. The control experiment DDPM is identical but does not contain the LD-DCMAC block. The baseline is a simple U-net.

exact distribution of the latent distribution and to apply corrections if not [180, 170]. We refer to this step as Langevin Diffused DCMAC (LD-DCMAC).

It should be noted that whilst we use a conditional model $p_k(\mathbf{x_k}|\mathbf{y})$ in our study, it is also possible to use an unconditioned model. The key part that changes is that our reverse noising process needs to be modified to reflect that we no longer wish to take a random walk across the entire data distribution but instead remain within some locally vicinity of the latent representation. This is the approach also used by [194, 180] where they find the motivation from attempting to perform something that resembles a proximal mapping operation during each reverse diffusion step.

The full implementation of our proposal is outlined in Algorithm 1.

6.3.1 Motion Estimates

The optimisation in section 6.1.2 assumes a jointly training motion estimator. However, the training scheme for the DDPM as proposed does not natively incorporate this capability. Instead, we train our motion estimator separately using the identical autoencoder-like network from section 3.7. The choice of this network is motivated by the aim to demonstrate that our proposed method does not require high quality motion estimates. Different datasets will have different types of motion thus it may be harder to perform motion estimation and motion estimates may not be unique with variations from one network initialisation to another [37].

By using this crude motion estimator, we hope to demonstrate the robustness of our motion-exploiting DDPMs even when motion estimates are suboptimal. For example, out-of-training-domain cases for particular cardiac views may generate optical flows with too much or too little displacement. Just as in section 3.7, higher quality motion estimates can be used in future use of our proposed method.

6.3.2 Score Function

We train the score function using the same methodology as in [161]. This involves making use of the exact distribution at each diffusion step being known in closed form due to the choice of drift and volatility terms in the SDE that set up the diffusion model. At each training step, a random diffusion step between $k = \{0...K\}$ is uniformly chosen. A point within the set of latent representations at k for the training example is drawn, $\mathbf{x_k}$ - this involves a weighted sum of the training example, $\mathbf{x}$, and normal noise, $\mathbf{z}$, as in equation (6.10).[3] The $\mathbf{x_k}$ term and k is provided to a U-net which is trained to predict the noise present in $\mathbf{x_k}$ according to the loss function in equation (6.11).

$$\mathbf{x_k} = \sqrt{\tilde{\alpha}}\mathbf{x} + \sqrt{1-\tilde{\alpha}}\mathbf{z} \tag{6.10}$$

[3] In the case of variance preserving models, the variance at any diffusion step is aimed to be constant and independent of k

$$||\epsilon_\theta(\mathbf{x}, k) - \mathbf{z}||_2^2, \tag{6.11}$$

where ϵ_θ is the U-net which is trained to learn the score function of the data distribution at every diffusion step k.

6.3.3 Architecture and Dataset

The U-net used in the study is the same as from [184] except uses 3D convolutions (for temporal dimension) and we used a channel multiplier of 16 rather 32 to ensure we can fit the network into GPU memory. This U-net consists of 5.5M parameters[4]. The methodology of this study is summarised in Figure 6.1.

We used cardiac cine data from the UK BioBank study with over 24,000 scans, as in chapter 4. These are magnitude only images and hence we generate synthetic phases using the approach from [116] in order to disrupt any k-space symmetry that may occur during retrospective undersampling.

We trained our PyTorch implementation with a batch size of 1 on the full 2D cine slice; patches were not used. Training took 10 days on a 48GB NVIDIA RTX A6000 GPU but we speculate that this training time can be significantly reduced by increasing the batch size by using a patch based training scheme [93, 141] or more GPUs in parallelised training.

[4]Implementation of this U-Net can be found at `https://github.com/openai/guided-diffusion`

Algorithm 1 MEDDPM Inference using DCMAC and DCLD (Data Consistent Langevin Diffusion)

Require: T, number of temporal frames

Require: K, number of diffusion steps

Require: K_{limit}, the diffusion step at which DCMAC iterations stop being applied

Require: M_t, the motion estimate from the t'th frame to $t+1$'th frame $\forall t \in \{1...T\}$

Require: N_{DCMAC}, number of DCMAC iterations

Require: y_t, D_t for $t = 1...T$

$\mathbf{x_K} \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$

for $k = K...1$ **do**

$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$ if $k > 1$ else $\mathbf{z} = \mathbf{0}$

$\mathbf{x'_k} \leftarrow \frac{1}{\sqrt{\alpha_k}}\left(\mathbf{x_k} - \frac{1-\alpha_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(\mathbf{x}_k, k)\right) + \sigma_k \mathbf{z}$ ▷ Reverse Diffusion Step

if $k > K_{\text{limit}}$ **then**

for $i = 1...N_{\text{DCMAC}}$ **do** ▷ LD-DCMAC Step

$\mathbf{x'_k} \leftarrow \mathbf{M}\mathbf{x'_k}$ ▷ Motion Warp

$\mathbf{x'_k} \leftarrow \text{Shift}(\mathbf{x'_k}, 1)$ ▷ torch.roll (circular shift)

$x_k \leftarrow \text{DCLD}(\mathbf{x_k}, K_L)$

end for

end if

$\mathbf{x_k} \leftarrow \text{DCLD}(\mathbf{x'_k}, K_L)$ if $k > 1$ else $\text{DC}(\mathbf{x'_k}, \mathbf{y})$

end for

return $\mathbf{x}_0$

Algorithm 2 DCLD (Data Consistent Langevin Diffusion)

Require: x_k estimate of latent variable

Require: K_L, number of Langevin diffusion steps

Require: k, current diffusion step

Require: y acquired data for data consistency

$\mathbf{x_k^0} \leftarrow \mathbf{x_k}$

for $\tau = 1...K_L$ **do**

$\mathbf{x_k^{\tau\prime}} \leftarrow DC(\mathbf{x_k^{\tau-1}}, \sqrt{\bar{\alpha}_k}\mathbf{y})$ ▷ Latent Data Consistency Step

$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$ ▷ Generate Random Diffusion Noise

$\mathbf{x_k^\tau} \leftarrow \mathbf{x_k^{\tau\prime}} - \delta_k \mathbf{s}_\theta(\mathbf{x_k^\tau}, k) + \sqrt{2\delta_k}\mathbf{z}$ ▷ Langevin Diffusion Step

end for

return $\mathbf{x_k^\tau}$

6.4 Results

We evaluate our proposed method with an aggressive acceleration rate of x16 with two suitable baselines. We used 8 central frequency-encoding (FE) lines and a Cartesian Gaussian-distributed variable density sampling mask. The first vanilla U-net trained with an L2 loss. The second is a data guided DDPM as presented in [194, 180] in the case of noiseless data distributions ($\lambda \to \infty$). For our proposed ME-DDPM, we use $K = 1000$ diffusion steps, $N_{\text{DCMAC}} = 1$ DCMAC step per diffusion step and $K_L = 1$ Langevin diffusion step per DCMAC step. We do not apply DCMAC steps for the last $K_{\text{limit}} = 50$ diffusion steps to reduce the impact of imperfect motion fields being used in the reconstruction process. For the Langevin diffusion step size, ϵ_k we use:

$$\delta_k = \frac{(\gamma\sigma_k)^2}{2}, \tag{6.12}$$

where we found $\gamma = 0.1$ to generate satisfactory reconstructions. We use 3-fold cross validation to evaluate the performance of our models. For the test fold we used a smaller subset of 600 images for evaluation.

Some example reconstructions are shown in Figures 6.2-6.5 and a quantitative evaluation in Table 6.1.

Table 6.1: Results of dynamic reconstruction with x16 undersampled data. The MEDDPM provides a vast enhancement over the DDPM from [194, 180].

Model	PSNR	SSIM
Baseline U-net	28.33 ± 0.76	0.834 ± 0.060
DDPM (SOTA) [194, 180]	30.72 ± 1.07	0.913 ± 0.015
MEDDPM (Ours)	$\mathbf{34.33 \pm 1.07}$	$\mathbf{0.962 \pm 0.010}$
Difference		
DDPM - Baseline	2.39 ± 0.92	0.079 ± 0.059
MEDDPM - Baseline	6.00 ± 0.80	0.128 ± 0.058
MEDDPM - DDPM	3.61 ± 0.50	0.049 ± 0.009

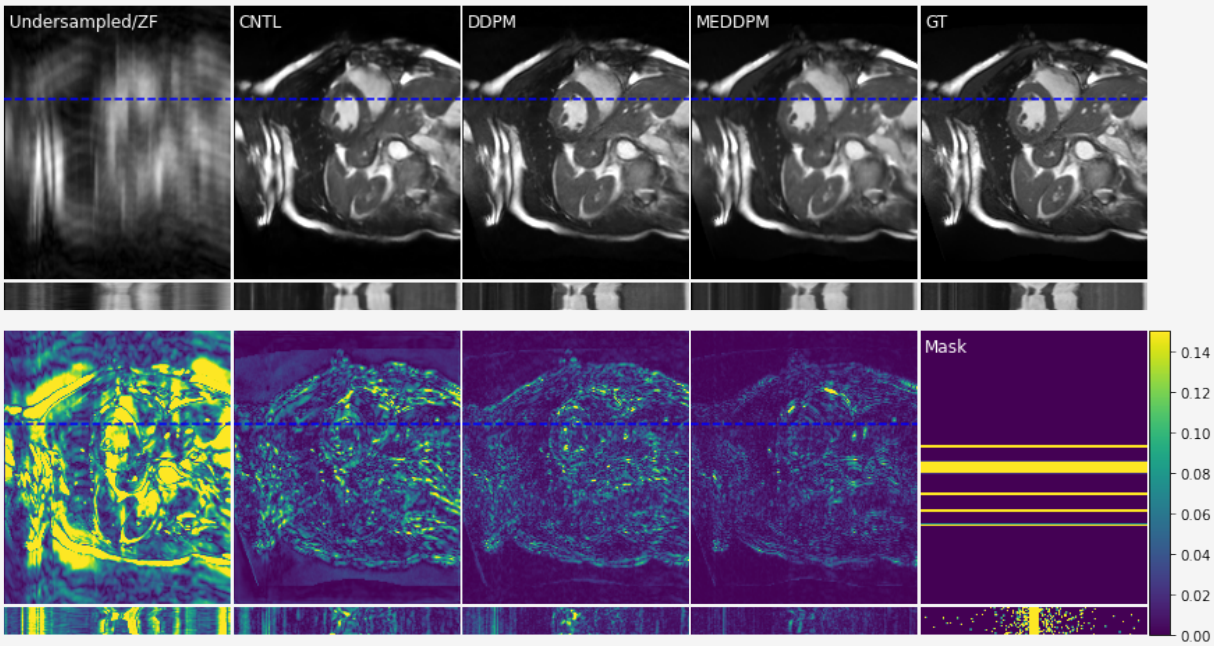

Figure 6.2: Reconstruction outputs from baselines and our proposed model alongside the ground truth. The (PSNR, SSIM) of the CNTL (Baseline U-net), DDPM and MEDDPM are: (28.0, 0.804), (30.1, 0.885), (33.5, 0.945).

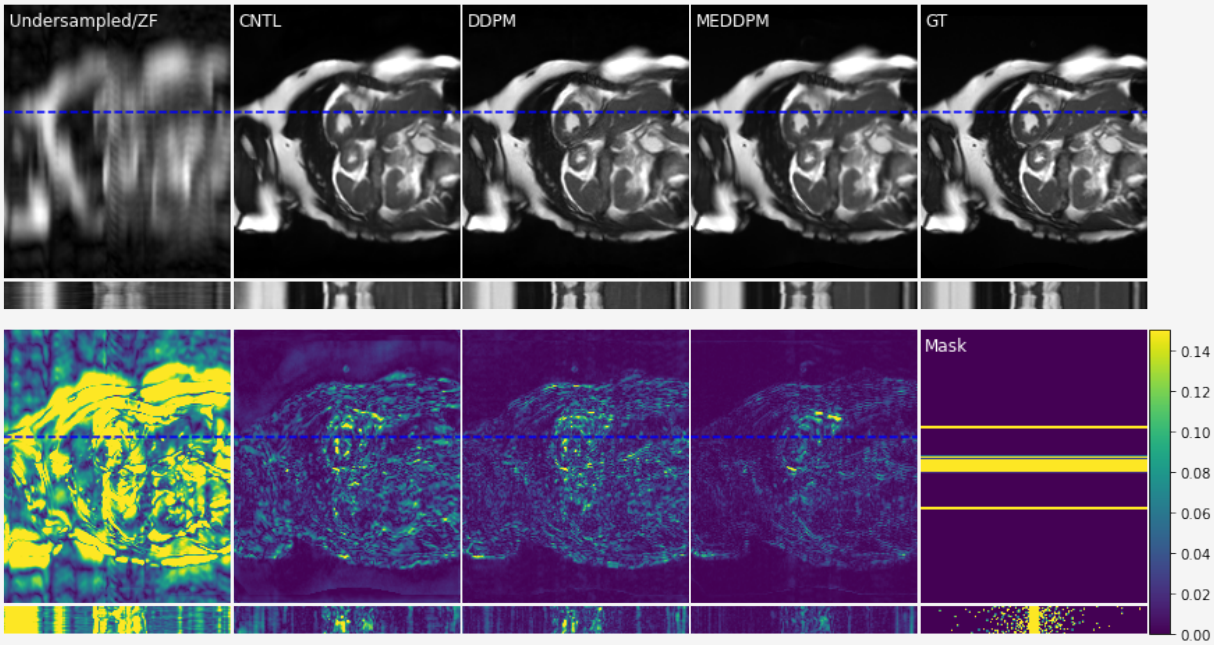

Figure 6.3: Reconstruction outputs from baselines and our proposed model alongside the ground truth. The (PSNR, SSIM) of the CNTL (Baseline U-net), DDPM and MEDDPM are: (29.6, 0.884), (31.2, 0.936), (35.3, 0.973).

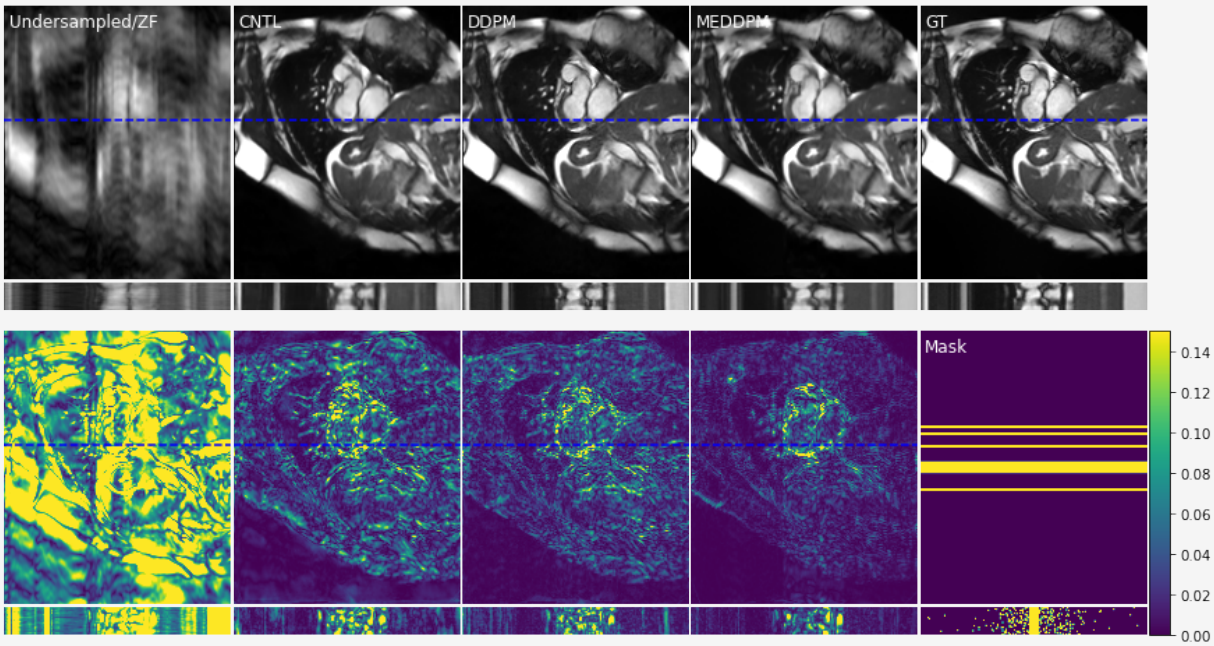

Figure 6.4: Reconstruction outputs from baselines and our proposed model alongside the ground truth. The (PSNR, SSIM) of the CNTL (Baseline U-net), DDPM and MEDDPM are: (27.1, 0.864), (28.9, 0.907), (32.2, 0.957).

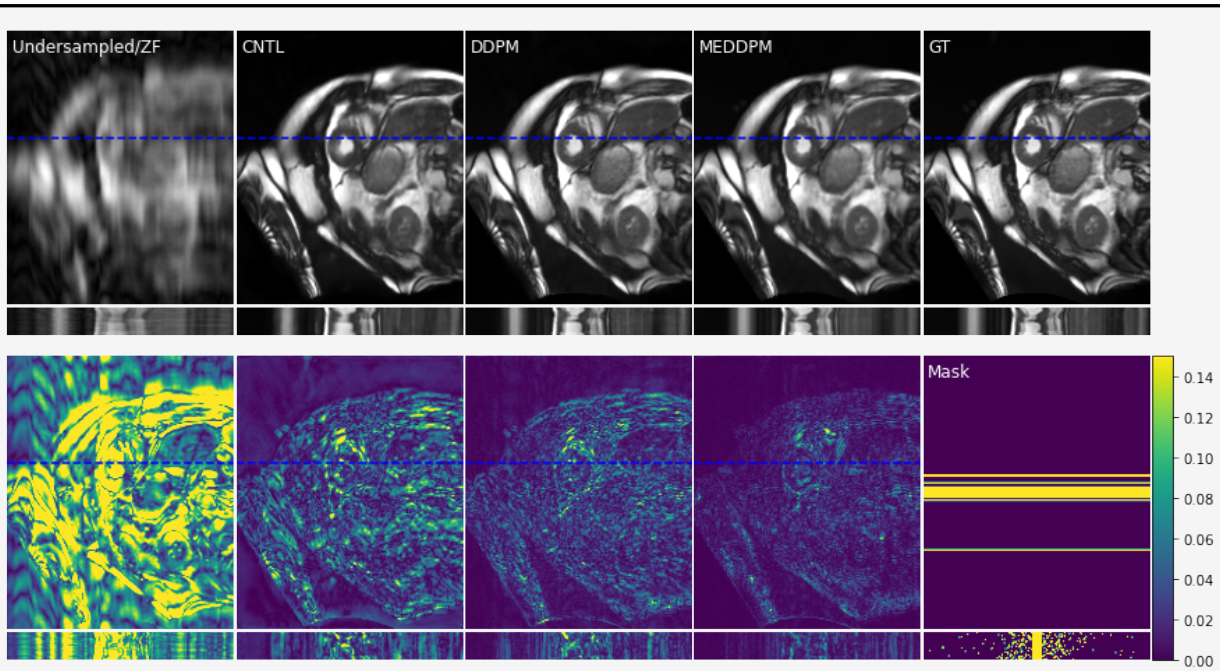

Figure 6.5: Reconstruction outputs from baselines and our proposed model alongside the ground truth. The (PSNR, SSIM) of the CNTL (Baseline U-net), DDPM and MEDDPM are: (27.5, 0.834), (30.1, 0.917), (34.5, 0.967).

6.5 Discussion

As shown in the results, the ME-DDPM out-performs the standard diffusion model and the baseline U-Net. Our ME-DDPM model for dynamic MRI reconstruction is able to leverage the benefit of motion via the DCMAC steps whilst mitigating the potential artefacts introduced by the DCMAC step. Once DCMAC steps are applied, our model uses the gradient of the learned data distribution to perform gradient descent towards the nearest point that matches our motion augmented latent representation.

This results in performance gain of 8.4% of the ME-DDPM over the DDPM compared with a vanilla U-net model. A Wilcoxon signed-rank test confirmed the performance of the ME-DDPM with $p \ll 0.01$. It should be noted that the motion estimates used in the DCMAC step were far from optimal - there is a vast literature on motion estimate/registration that produce estimates of much higher quality (see section 3.7 for more information). These estimates would likely increase the reconstruction quality even further.

We note that training the ME-DDPM was much more straightforward compared to GANs and comes with the benefit of higher fidelity reconstructions. Furthermore, it was surprising that our model was trainable with only a batch size of 1 - GANs typically do not work well in this regime.

Due to the probabilistic capabilities of the diffusion models, the ME-DDPM should be able to better adapt to data corruption scenarios such as motion corruption. We leave this for future work. Further to this, it is unclear whether DDPMs can adapt to different domains as friendly as other vanilla, discriminative reconstructions models (see Chapters 3 and 4; [134, 141]). Domain adaptation is an important topic in MRI reconstruction and more generally, in deep learning and thus warrants significant study in future advancement of this work.

In order to sample from the ME-DDPM, it requires $K = 1000$ forward passes through a computational intensive U-net which took 10 minutes on an NVIDIA RTX A5000 with 24GB of memory. This removes the possibility of real time imaging with our proposed method.

However, it should be noted that the probabilistic capabilities of DDPMs may allow the ME-DDPM to be leveraged when training with highly corrupted training data e.g. motion corrupted acquisitions (see section 7.5).

6.6 Conclusion

One of the main limitations of this study is that training DDPMs bares a large computational cost which restricted us to a reduced capacity U-net model. The inference time is also large as with all diffusion models. However, there exists work that focuses on reducing this inference time constraint [181]. Future work should investigate incorporating such speed increases with the ME-DDPM model.

Training the ME-DDPM is almost as straightforward as training conventional deep learning models making a convenient solution for the problem of motion-based reconstruction. Higher fidelity images whilst being able to model the general statistics of the training distribution are desirable properties in the future of MRI reconstruction. However, there are a few unanswered questions that are of utmost importance, as mentioned above. Overall, ME-DDPMs offer a promising direction of research in dynamic MR reconstruction.

Chapter 7

Unsupervised MRI reconstruction with Diffusion Models

Diffusion Models are a class of generative models that have shown to produce high-fidelity results in the case of accelerated Magnetic Resonance (MR) image reconstruction. In MR imaging, there are a wide range of imaging scenarios where obtaining fully sampled k-space data is challenging which results in limitations in spatial and temporal resolution. Recent work shows that MR reconstruction can be performed with deep learning without the need of fully sampled data during learning. In this paper, we show that diffusion models can be used to mitigate a variety of imaging challenges that MR reconstruction presents. We propose how diffusion models can be used to reconstruct accelerated acquisitions when the training targets are noisy or not fully sampled. The lack of ground truth data presents an issue that diffusion models can overcome. We demonstrate this with studies using cardiac data from the UK BioBank study and knee MR images from the fastMRI challenge. We also show that diffusion models outperform a variety of baselines with over 10% increases in PSNR and SSIM in some scenarios.

7.1 Introduction

There are many scenarios whereby full acquisitions of k-space data required for the reconstruction of MR images are made difficult by issues such as respiratory motion, patient discomfort and slow scan times. This results in limitations in spatial and temporal resolution. To mitigate the impact of such issues, methods are sought to accelerate the acquisition of k-space data. Accelerated acquisition usually involve the undersampling of k-space and parallel imaging can also be used.

The problem of accelerated MR image reconstruction stems from that data acquisition takes place in a Fourier space, the so-called 'k-space', rather than in image space. A condition for reconstruction from Fourier coefficients is that one must sample enough data such that we do not violate the Nyquist sampling criterion [40]. When this criterion is violated, compressed sensing formulates a method by which the true image can still be recovered [40]. The specific optimisation problem by which image recovery occurs is detailed by equation (7.1).

$$\frac{\alpha}{2}||D\mathcal{F}x - y||_2^2 + \mathcal{R}(x), \tag{7.1}$$

where D is an undersampling mask that represents the acquired points in k-space, x is the reconstructed image, y is the acquired single-coil k-space data, $\mathcal{F}$ is the Fourier transform and $\mathcal{R}$ is the image regularisation. There is a vast literature on choosing and/or learning a regulariser to perform the above optimisation [93, 140, 110, 16, 188, 36, 198], especially in a supervised fashion where fully-sampled data exists for use in the training objective. In this paper we primarily focus on the scenario where fully-sampled k-space data does not exist and thus we must operate in the unsupervised or self-supervised setting. We conduct preliminary experiments on the assumption of single-coil data but in section 7.6 we also investigate multi-coil data.

Unrolled reconstruction networks Whilst classical reconstruction methods such as FISTA do not rely on fully sampled data, they are typically restricted to less aggressive acceleration

factors and low image fidelity [41]. One possible iterative decomposition of the optimisation in equation (7.1) is through equations (7.2) and (7.3) where the former represents a proximal mapping learned by convolutional neural network (CNN) and the latter is a data consistency step to hold the reconstruction true to the acquired data [93, 123, 49]:

$$u^k = \text{CNN}_k(x^{k-1}) \tag{7.2}$$

$$x^k = (\lambda \mathcal{F}^T D^T D \mathcal{F} + I)^{-1}(\lambda \mathcal{F}^T D^T y + u^k), \tag{7.3}$$

Here x^k is the estimated reconstruction at iteration k and λ controls the level of data consistency ($\lambda \to \infty$ in the noiseless case). This is similar to the approach of the DC-CNN [93, 123] rather than variational networks which are gradient descent based [110].

Loss functions and data corruption A popular choice of loss function for MRI reconstruction is the L2 loss or variation thereof [140, 126, 177]. Hybrid losses have also been used whereby GAN-like losses are used to create (medically) photorealistic outputs [98]. The motivation behind using an L2 loss is for maximising the likelihood of the model output (see section 2.4).

In practice, for most types of high quality, noiseless data, an L2 loss works sufficiently well as shown in previous literature e.g. [110, 93, 177, 121]. However, a fundamental assumption is that the distribution of model errors is identically normally distributed. For MRI data, this may not necessarily be true particularly in case of motion corruption. When imaging targets in supervised learning are magnitude images, the noise is also non-Gaussian (but instead Rician). In the case of a Rician distributed noise such as with magnitude-only images, an L2 loss would not be directly maximising the likelihood and hence would result in a suboptimal optimisation procedure.

However, our targets may not necessarily be magnitude images but rather complex-valued images which contain a noise that is much closer to being Gaussian rather than Rician. Whilst it may appear that an L2 loss would be appropriate under this scenario, another assumption

of using an L2 loss is that the noise in the imaging target is identically and independently distributed (i.i.d.) across the dataset. However, this is not necessarily true in the case of certain datasets such as the fastMRI dataset where images come from different scanners. Additionally, in the multicoil case, some coils may contain more noise than others depending on the imaging subject and scanner geometry [158, 12, 145].

Aims and Contributions We introduce the use of score-based generative models for self-supervised MR reconstruction without fully sampled data. In particular, we use a denoising diffusion probabilistic model (DDPM) to efficiently train a neural approximation to our required data density [161, 170]. The trained DDPM model substitutes the CNN in equation (7.2). The main issue of using DDPMs to solve this problem is that we lack knowledge of the fully-sampled latent distribution. In our study, we propose a solution to this problem. Our main contribution is proposing how to use DDPMs in a self-supervised setting in the context of MRI reconstruction. In particular, we focus on *diffusion models for unrolled Cartesian accelerated MR reconstruction without fully sampled data and only noisy acquisitions.* The main motivations behind the use of DDPMs to solve this problem are as follows:

1. These generative models provide an easy way to sample from the approximate conditional distributions whilst also being straightforward to train unlike conventional GANs. In the case of very corrupted k-space data such as noisy data from low-field MRI scanners or motion during the acquisition, generative models can be leveraged to mitigate for this data corruption in the reconstruction process

2. DDPMs have been shown to provide high fidelity results competing with vanilla baselines [194, 180]. It's possible they may perform competitively in the case of self-supervised MRI reconstruction

3. DDPMs perform (approximately [193, 161, 182]) maximum likelihood training without requiring prior knowledge of the noise distribution of the data

The main contributions of our study are as follows:

1. A general proof of concept that diffusion models can be used in the unsupervised scenario to perform MRI reconstruction which has not been previously studied;

2. A direct way to extend DL-MRI reconstruction to multiple reconstruction steps (not diffusion steps) whilst still leveraging the generative advantages of diffusion models. Incorporating the concept of network unrolling with DDPMs helps generate high fidelity reconstructions compared to previous studies such as [180, 194];

3. Show that diffusion models can be directly leveraged for mitigating the issue of training with noisy MRI data;

4. Present an example of noise-mitigated MRI reconstruction without fully sampled training data from the fastMRI single-coil and multi-coil reconstruction challenge.

The structure of the main content in this chapter is summarised below:

1. Section 7.3 — MRI reconstruction without Fully Sampled training targets using our proposed 1-DDPM ('unsupervised'/self-supervised learning);

2. Section 7.4 — Extension of DDPM-based MRI reconstruction to multiple reconstruction steps (not diffusion steps) with our proposed score decomposition (with and without fully-sampled training targets);

3. Section 7.5 — Exploration of the use case of DDPMs for mitigating the impact of noisy imaging targets (with fully-sampled training targets);

4. Section 7.6 — Using the above three sections for noise-mitigated, unsupervised cascaded MRI reconstruction using our proposed DC2DDPM as demonstrated on the fastMRI knee dataset.

7.2 Data

The dataset used in our study consists of 1,000 3D multi-slice short-axis cardiac cines taken from the UK BioBank study[1] [67]. We extract all apical, mid-cavity and basal slices and convert the MR cine into a stack of 2D magnitude images cropped to a size of 192 x 192. We normalise that data between -1 and 1 using the 99th percentiles within each short-axis volume. We used 700 of the cines during training and test the methods in this study on 1000 images sampled randomly from the remaining 300 cines. Associated with each extracted 2D image in our dataset, there is also a Cartesian undersampling mask with a Gaussian distribution and x4 acceleration rate (25% sampling).

To conclude the chapter, we evaluate our method on the fastMRI knee dataset. The details of this dataset are given in section 7.6.

7.3 Generative Models for Unsupervised MRI Reconstruction with DDPMs

7.3.1 Related works

Unsupervised MRI reconstruction There have been a range of studies to perform MRI reconstruction without complete k-space data [36, 41, 177, 154]. In the study by [177], they introduce a k-space sampling strategy that results in high fidelity reconstructions whilst lacking fully sampled data at training time. The idea is that the training set consists of k-space images each acquired with its own undersampling mask. At training time, the acquired lines in the undersampling mask, $\mathbf{\Omega}$, are divided into two disjoint sets, $\mathbf{\Lambda}$ and $\mathbf{\Omega}/\mathbf{\Lambda}$. The strategy for choosing $\mathbf{\Lambda}$ is to choose lines in $\mathbf{\Omega}$ according to a Gaussian distribution. The fraction of the number of lines in this subset, ρ, which results in highest fidelity reconstructions was found to be $\rho = 0.4$ [177]. The network architecture used resembles that of an unrolled optimisation

[1] `http://imaging.ukbiobank.ac.uk`

whereby each CNN block represents a proximal operation (equation (7.1)) and is immediately followed by a data consistency step (equation (7.3)) [93, 177]. In the case of [177], the input to the network is data at the locations in mask $\mathbf{\Omega}/\mathbf{\Lambda}$ which is also used in the data consistency steps. However, the training loss is only computed on the k-space lines in the $\mathbf{\Lambda}$ mask. This allows the network to implicitly learn the full reconstruction whilst only being supervised to reconstruct part of it.

Generative adversarial networks (GANs) have provided inspiration to the problem of MR reconstruction in various forms [188, 154, 102]. In particular, [154] use the adversarial training scheme of WGANs to also implicitly perform MR image reconstruction without fully-sampled data. A generator which takes the form of an unrolled optimisation network produces an output given a zero-filled reconstructed acquisition. This output is subsequently undersampled with a random mask that is different to that of the input/training data. A discriminator is then trained to discriminate between real, undersampled data and these undersampled outputs. This discriminator is subsequently used to train the generator to produce better samples which the generator can only do from implicitly learning to reconstruct the fully sampled image from the acquired/undersampled data.

Generative Modelling with Score-based Models In terms of generative modelling, recent work in the field has focused on the use of stochastic differential equations (SDEs) for approximating data distributions [170, 193]. This relies on mapping the data distribution $q(\mathbf{x})$ to $N(\mathbf{0}, \mathbf{I})$ through a forward diffusion process denoted in equation (7.4):

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w} \tag{7.4}$$

Here x is our latent variable being diffused, w is a Wiener process and t is the index of said process. Interestingly, the reverse process of equation (7.4) is also a diffusion process and can be written as equation (7.5) [1, 170]:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_x \log q_t(\mathbf{x})]d\tilde{t} + g(t)d\tilde{\mathbf{w}}, \tag{7.5}$$

Here $\tilde{w}$ and $\tilde{t}$ are in the reverse time direction and $q_t(\mathbf{x})$ is the distribution of the diffusion of our data at diffusion step t. The reverse diffusion process requires the gradient of the log data density which can be off-loaded to a conditioned neural network giving us a score function, $\mathbf{s}(\mathbf{x}, t)$. [194, 180] show that data consistency can be used in the reverse diffusion process to guide the latent representation towards performing undersampled MRI reconstruction.

In the case of DDPMs, the drift and volatility terms are set by the functions in equations (7.6) and (7.7) for $\beta_t << 1$. This is similar to the diffusion process in variance preserving SDEs [170]. This choice allows the calculation of any latent sample x_t given x_0 which results in an efficient and simple training process [161]:

$$\mathbf{f}(\mathbf{x}, t) = -\frac{1}{2}\beta_t \mathbf{x} \tag{7.6}$$

$$g(t)^2 = \beta_t \tag{7.7}$$

As in [161], we define some useful quantities that control the stochastic process: $\tilde{\alpha}_t = \prod_t (1-\beta_t)$ and $\beta_t = \beta_0 + \frac{(\beta_T - \beta_0)}{T} t$ where T is the chosen number of diffusion steps in the model and $\beta_0, \beta_T \ll 1$ controls the prescribed variance schedule. In DDPMs, instead of using a weighted sum of score matching losses to directly maximise the log-likelihood [193], an unweighted L2 loss is used that predicts the amount of noise present in each latent variable, $\mathbf{x}_t$:

$$L = \mathbb{E}_{\mathbf{x}_0, \epsilon, t}\Big[||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||^2\Big], \tag{7.8}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, θ are the parameters of the neural network and the following relation between the score and ϵ_θ exists:

$$s_\theta(\mathbf{x}_t, t) = -\frac{\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{1 - \tilde{\alpha}_t}}. \tag{7.9}$$

7.3.2 Method

With our undersampled data, it is not possible to use DDPMs in the traditional form to approximate the fully-sampled data densities $\nabla q_t(x)$. This is simply because without knowledge of the ground-truth data, we cannot simulate the forward diffusion process. Instead, we condition our data density (and thus our learned score function) on the k-space line number(s) that we are trying to predict which we denote as $\mathbf{\Lambda}$. Our score function is trying to learn $\nabla q_t(x_{\mathbf{\Lambda}}|\mathbf{\Lambda})$. In other words, the score function only operates on the k-space lines contained within $\mathbf{\Lambda}$. One possible choice for $\mathbf{\Lambda}$ is that it contains only a single line in k-space.

This conditioning is possible in a way that allows training to take place efficiently. This is due to how k-space frequencies in Cartesian undersampling present themselves in image space. A particular point in Fourier space generates a continuous sinusoid in the image domain with a specific number of peaks, troughs and phase. In the forward diffusion process, we only add noise to the line(s) $\mathbf{\Lambda}$ in k-space. This process is summarised in the equation below:

$$x_{t,\mathbf{\Lambda}} = \mathcal{F}^T D_{\mathbf{\Lambda}} \mathcal{F}(\sqrt{\tilde{\alpha}_t}\mathcal{F}^T D_{\mathbf{\Lambda}}^T y + \sqrt{1-\tilde{\alpha}_t}\epsilon). \tag{7.10}$$

This makes it possible for a CNN to better condition itself to predict information for a particular line in k-space whilst still operating in the image domain. The k-space position(s) is implicitly encoded rather than directly encoded like with positional encoding found in Transformers [95].

DDPM conditioning Along with the $\mathbf{\Lambda}$ condition, we also condition our score function on the acquired data. During inference, this conditioning consists of all of the acquired k-space data, $y = y_\Omega$. During training, this condition is changed to $y_{\Omega/\Lambda}$ so that it does not include data that the score function is trying to predict. This takes inspiration from the work in [177] where they train with a loss function only on Λ to implicitly learn to predict the entire underlying image. In our study, we cannot leverage exactly the same idea since we require a known latent distribution at training time which forces us to explicitly learn to score Λ only. We study two settings for Λ: $\rho = 0.03$ which is equivalent to having a single k-space line in Λ and $\rho = 0.4$

which was found to be optimal in the case of CNNs in [177].

By leveraging k-space based positional encoding, we present our proposed method in algorithms 1 and 2 for training and inference respectively. The training of the model occurs in the same style as with DDPMs [161] except the latent variables only contain the lines Λ in the latent k-space. Inference is also identical to DDPMs except we generate reconstructions for only the lines Λ and hence multiple inference passes are required with different Λ to ensure each line in the k-space is reconstructed at least once. All code was implemented in PyTorch and will be available on GitHub.

Algorithm 3 Training

while not converged **do**

$\mathbf{x_{0,\Omega}} \sim \mathbf{q}(\mathbf{x_{0,\Omega}})$

$t \sim \text{Uniform}(\{1...T\})$

$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$

$\mathbf{\Lambda} \sim \text{Gaussian}(\mathbf{\Omega}; \rho)$

$\epsilon_{\mathbf{\Lambda}} = \mathcal{F}^T D_{\mathbf{\Lambda}} \mathcal{F} \epsilon$

$\mathbf{x_{0,\Lambda}} = \mathcal{F}^T D_{\mathbf{\Lambda}} \mathcal{F} \mathbf{x_{0,\Omega}}$

$\mathbf{x_{\Omega/\Lambda}} = \mathcal{F}^T D_{\mathbf{\Omega/\Lambda}} \mathcal{F} \mathbf{x_{0,\Omega}}$

$\mathbf{x_{t,\Lambda}} = \sqrt{\tilde{\alpha}_t} \mathbf{x_{0,\Lambda}} + \sqrt{1 - \tilde{\alpha}_t} \epsilon_{\mathbf{\Lambda}}$

$\hat{\epsilon}_{\mathbf{\Lambda}} = \epsilon_\theta(\mathbf{x_{t,\Lambda}}, t | \mathbf{x_{\Omega/\Lambda}})$

Gradient Descent: $\nabla ||\hat{\epsilon}_{\mathbf{\Lambda}} - \epsilon_{\mathbf{\Lambda}}||^2$

end while

Algorithm 4 Inference

Require: $\mathbf{\Lambda}_1 ... \mathbf{\Lambda}_N s.t. \mathbf{\Lambda_i} \notin \mathbf{\Omega}$

Require: $\mathbf{\Lambda}_1 \cup \mathbf{\Lambda}_2 \cup ... \cup \mathbf{\Lambda}_N \cup \mathbf{\Omega}$ samples every k-space line at least once

Require: acquired data, $\mathbf{x_\Omega}$, $\mathbf{D_\Omega}$

$\mathbf{x_T} \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$

$\mathbf{D}_{\tilde{\mathbf{\Lambda}}} = \sum_i \mathbf{D}_{\mathbf{\Lambda_i}}$

for $t = T...1$ **do**

$\hat{\epsilon} \leftarrow \mathbf{0}$

$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathcal{I})$ if $t > 1$ else $\mathbf{z} = \mathbf{0}$

for $i = 1...N$ **do**

$\hat{\mathbf{x}_{t,\mathbf{\Lambda_i}}} \leftarrow \mathcal{F}^T \mathbf{D}_{\mathbf{\Lambda_i}} \mathcal{F} \mathbf{x_t}$

$\hat{\epsilon} \leftarrow \hat{\epsilon} + \mathbf{D}_{\mathbf{\Lambda_i}} \mathcal{F} \epsilon_\theta(\hat{\mathbf{x}}_{\mathbf{t,\Lambda_i}}, t | \mathbf{x_\Omega})$

end for

$\hat{\epsilon} \leftarrow \hat{\epsilon} \odot \frac{1}{\mathbf{D}_{\tilde{\mathbf{\Lambda}}}}$

$\hat{\epsilon} \leftarrow \mathcal{F}^T \hat{\epsilon}$

$\hat{\mathbf{s}} \leftarrow -\hat{\epsilon} / \sqrt{1 - \tilde{\alpha_t}}$

$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta}} \left(x_t + g^2(t) \hat{\mathbf{s}} \right) + g(t) \mathbf{z}$

end for

return $\mathbf{x}_0 + \mathbf{x_\Omega}$ (equation 3)

Experiment	PSNR	ROI-PSNR	SSIM	ROI-SSIM
CNTL (Yaman et al. 2019)	26.9 ± 1.3	25.8 ± 1.0	0.840 ± 0.039	0.895 ± 0.022
ρ-DDPM [Ours]	27.0 ± 1.3	25.0 ± 1.0	0.837 ± 0.035	0.892 ± 0.019
1-DDPM [Ours]	$\mathbf{29.5 \pm 1.7}$	$\mathbf{28.8 \pm 1.5}$	$\mathbf{0.867 \pm 0.037}$	$\mathbf{0.927 \pm 0.020}$
Difference:				
ρ-DDPM - CNTL	0.1 ± 0.4	0.2 ± 0.5	-0.002 ± 0.012	-0.002 ± 0.011
1-DDPM - CNTL	$\mathbf{2.6 \pm 0.7}$	$\mathbf{3.0 \pm 0.8}$	$\mathbf{0.028 \pm 0.016}$	$\mathbf{0.032 \pm 0.011}$

Table 7.1: Table of results using 1000 test images. The CNTL experiment refers to the model proposed in [177]. The PSNR and SSIM metrics are shown for the whole image as well as for a central crop of the image in the region of interest (ROI). Whilst faster inference can be obtained with ρ-DDPM, the image quality isn't as competitive as 1-DDPM. 1-DDPM outperforms the CTNL in both PSNR and SSIM. The last two rows of the table show the average (and standard deviation) of the difference in PSNR and SSIM calculated per example in the test set.

7.3.3 Experimental results

Unsupervised DDPMs In order to evaluate our proposed DDPM-based model, we train a Unet taken from [184] with 8.6M parameters that is conditioned on the diffusion index, t. As in [161], we use $T = 1000$ diffusion steps setting $\beta_0 = 0.0001$ and $\beta_T = 0.2$. We also train two versions of the model: one in which at training time, the score function only predicts a single line in k-space at a time, $\rho = 0.03$, (referred to as 1-DDPM), and one in which the score function predicts a small subset of k-space lines, $\rho = 0.4$ (referred to as ρ-DDPM). In the case of $\rho = 0.4$, we also performed a random search for a set of N masks $\mathbf{\Lambda_1}...\mathbf{\Lambda_N}$ such that every k-space line was fully-sampled at least once. We found that $N = 35$ was sufficiently large such that the random search finished within a few seconds. We also trained the model proposed in [177] with a single proximal mapping with data consistency for a fair comparison against our proposal. It should be noted that we used the same Unet structure here as in our proposed DDPM model with the same number of parameters [184]. The results are summarised in table 7.1. ROI refers to calculation of said metric in the central region of interest in the image rather than across the entire image. Some examples of the results are shown in figure 7.1 and some particular zoom-ins are shown in figure 7.2.

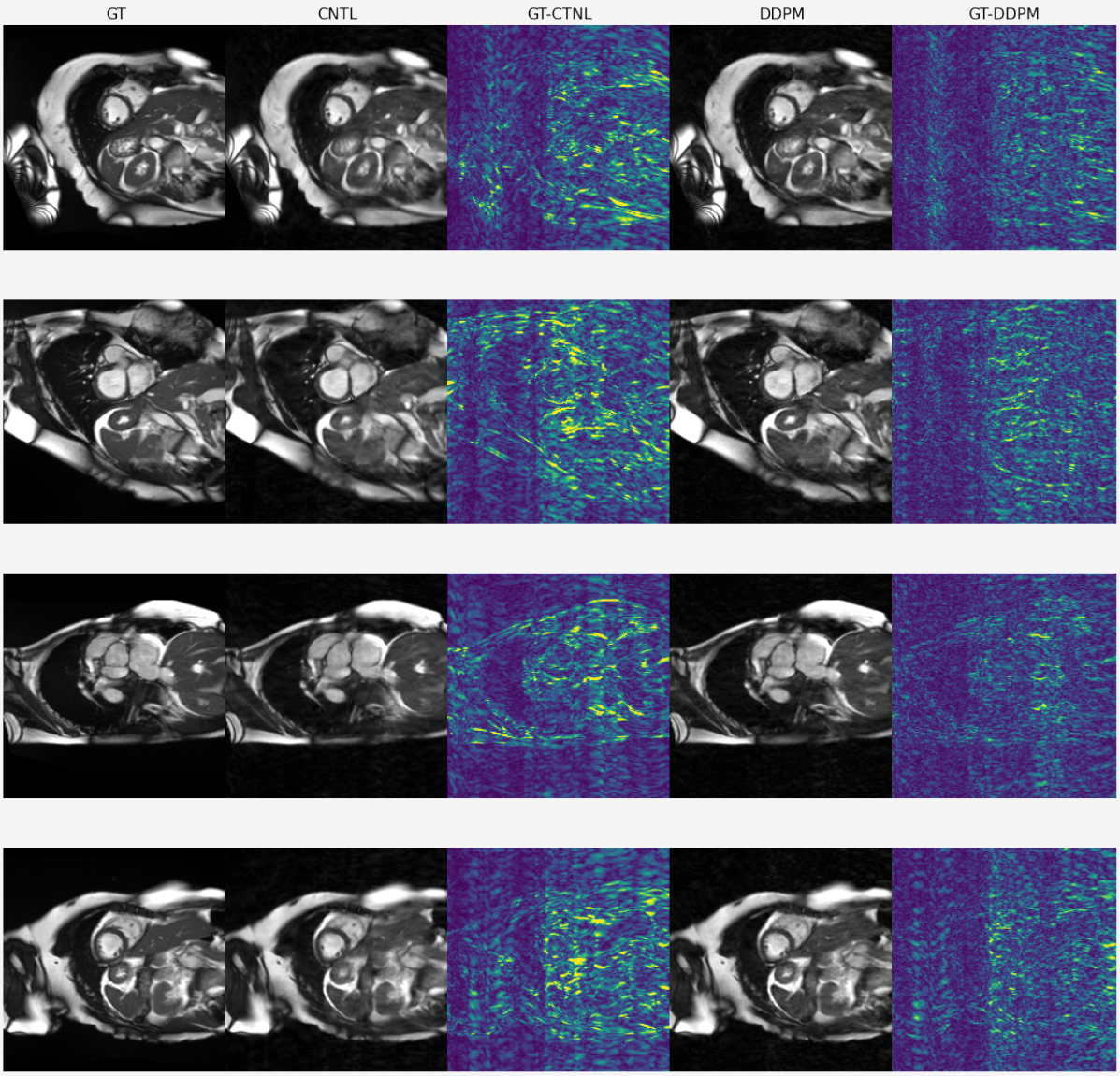

Figure 7.1: Some examples of results of 1-DDPM vs CNTL. Each row is a different example from our test set. "GT" refers to the ground truth image taken directly from the UK BioBank dataset before any retrospective undersampling. The third and fifth columns are the difference between the ground truth and the stated method. It can be seen that 1-DDPM generally produces fewer image artefacts compared to the CNTL experiment. In particular, edges in the 1-DDPM output are more sharply defined whilst also being less noisy overall.

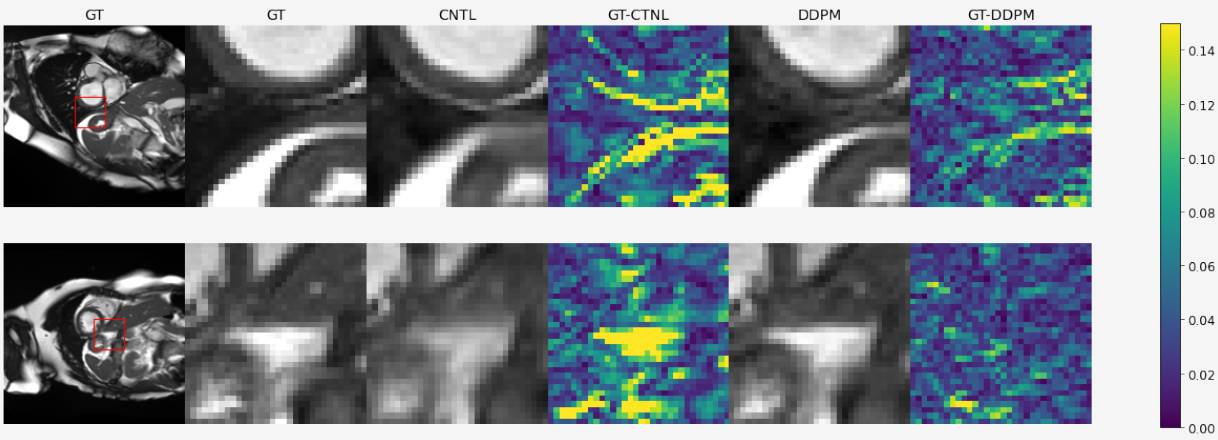

Figure 7.2: Results of 1-DDPM vs CNTL in paticular regions. We show some particular clear cases where the 1-DDPM model has produced a sharpness that the CNTL experiment hasn't been able to replicate.

7.3.4 Discussion

Unsupervised DDPM Our results show that DDPM models provide some performance gain when compared against the control experiment. Further to this, we found that using a single line $\mathbf{\Lambda}$ performs far better than with $\rho = 0.4$. This is likely due to the simplicity of being able to encode the spatial waveform introduced by single k-space line $\mathbf{\Lambda}$ (e.g. the network could implicitly perform an absolute or ReLU-based summation of the network input to determine the k-space line being predicted hence creating an implicit embedding for k-space position). For $\rho = 0.4$, after the mask $\mathbf{D}_{\mathbf{\Lambda}}$ is applied to the noise in Fourier space, the way in which positional data is encoded in the resulting masked noise in image space is much less obvious and thus harder for the network to learn and decipher.

Using a single NVIDIA RTX A6000 GPU, we found that a forward pass with our proposed method takes around 6 minutes for the single line $\mathbf{\Lambda}$. This could be parallelised with N GPUs to perform inference in around 30 seconds. However, the control experiment as presented in [177] takes only a few hundred milliseconds and thus presents a significant advantage for real-time imaging. In spite of this, our proposed method may still provide benefits for noisy or corrupted imaging scenarios and may have some use as a post-processing step.

7.3.5 Conclusion

In this study, we proposed using a DDPM instead of a vanilla CNN in a particular unrolling of the MRI reconstruction process. We conditioned our DDPM on the k-space lines that needed to be predicted. In particular, we studied the scenario where the DDPM predicted a single k-space line at a time, 1-DDPM, and the scenario where the DDPM predicted a small set of k-space lines at a time, ρ-DDPM. We found that both methods compete with the use of a vanilla CNN but in particular, the 1-DDPM significantly outperformed the vanilla CNN. In conclusion, this study has found that DDPMs present a performance gain in the problem of self-supervised MR reconstruction.

Whilst we've found that DDPM models can be used in place of a vanilla Unet such as that used in the control experiment of this study [177], the extension of this method to multiple reconstruction steps (not diffusion steps) remains an open research question. Extending the work of [177] (CNTL) to multiple reconstruction steps is straightforward and has been studied. In the following sections, we proceed to investigate how multiple iterative reconstruction steps can be incorporated with the DDPM and explain the limitations of current approaches.

7.4 Cascading reconstructions with DDPMs

[194] provide a method for incorporating DDPM guidance via manipulation of its latent space. In particular, they borrow the concept of data consistency from decomposed, unrolled MRI reconstruction with proximal operators. The closed form solution to the data consistency proximal step is used to force the latent representation to hold true to the acquired data. However, only the data consistency step is performed and without the context provided in an unrolled reconstruction scheme where the data consistency term naturally arises.

The diffusion model is using an iterative scheme to reverse a stochastic process at each step, not perform a direct gradient descent [110] or proximal operation towards the objective [93, 123, 177]. Instead, DDPMs can be viewed as a substitute for neural networks in unrolled

optimisation schemes but with added probabilistic capabilities (e.g. for modelling noise, random artifacts or motion corruption). This is opposed to the view that they are a new approach to decomposing MRI reconstruction.

Incorporation of the data consistency in the closed form solution given in [194] does not by itself mean that an iteration in a diffusion model is the equivalent of a proximal iteration or gradient descent in unrolled networks. The training objective of the DDPM can be written as performing a singular reconstruction step when conditioned on the undersampled data resulting in a score function that operates the same as when the data consistency from [194] is applied in an unconditioned setting (see section 7.4.1). The vanilla conditional DDPMs can be viewed as performing a single reconstruction step but with the ability to also model random noise in the target (which is off-loaded to the many diffusion steps in the model). This view can be summarised as DDPMs performing an *implicit reconstruction* to predict the artifical diffusion noise present in the latent variable (reconstruction step) whilst simultaneously *modelling the noise* in the target (probabilistic step).

Figure 7.3 provides an illustration depicting the difference between the iterations in guided DDPMs [194] and iterations used in unrolled schemes such as proximal-based reconstruction.

Another limitation in [194] is how data consistency is enforced. The work in [194] uses data consistency at each intermediate step however their score prediction is conditioned on the entire latent variable, including the parts where data has not been acquired and enforced (and there is no conditioning on the acquired data). This means that the data gradients can either descend to favour the data consistent parts or favour the parts where data does not exist.

Instead of proposing to perform MRI reconstruction solely of latent manipulation of DDPMs, we wish to introduce the use of unrolled networks prior to a DDPM stage. In this study, we choose a cascading network of data consistent Unets that perform a series of proximal operations with the last being performed by the DDPM. When conditioned on an iterative N-cascade unrolled NN, the end-to-end training of the DDPM with the conditioning NN simply adds a powerful, probabilistic proximal operator as the final step of the decomposition as explained in the following section.

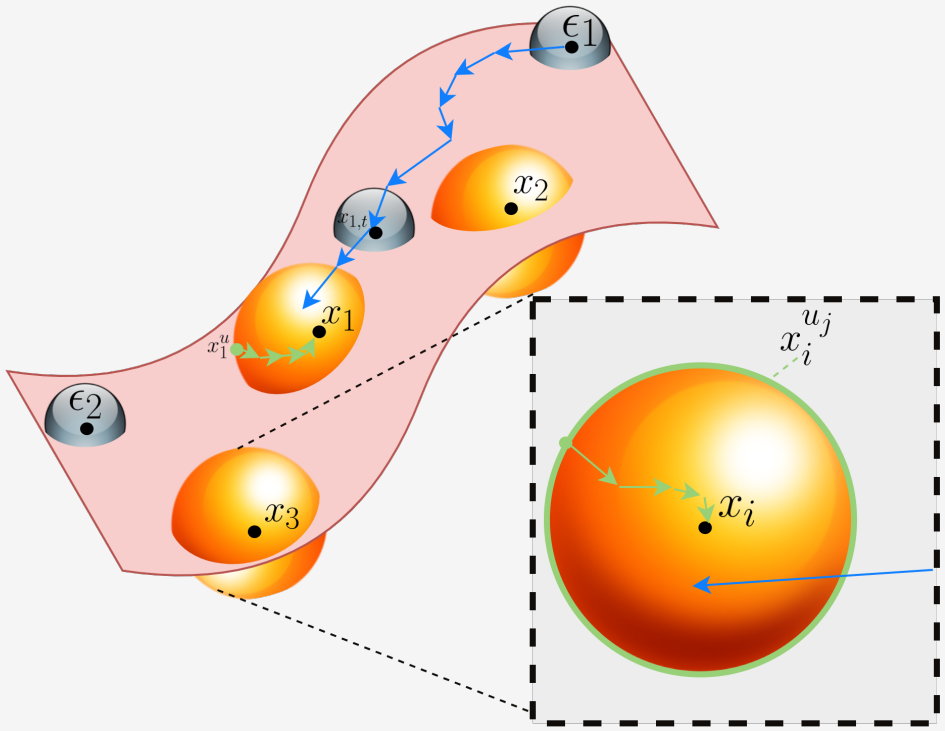

Figure 7.3: In this illustration black dots - centroids - represent the set of possible MRI images. This includes patient images as well as other images such as images of noise or phantoms. All centroids - MRI images - lie on a hyperplane depicted in red or grey. Surrounding these centroids are hyperspheres that are related to the centroid image, x. The surface of the hyperspheres represents the set of possible undersampled images, x^u. The volume in between the surface and centroid represent the set of possible reconstructions of x given that accelerated MR reconstruction is an ill-posed problem. The green dot represents a possible starting point for a traditional descent method to iteratively converge towards the centroid x. The centroids ϵ are the starting positions of guided-DDPM's reverse diffusion process ($\epsilon \sim \mathbb{N}(\mathbf{0}, \mathbb{I})$). The blue arrows represent the Euler-Maruyama descent to approximate the reverse diffusion process that guides it towards x. x_t represents a possible realisation of the t'th reverse diffusion step (of which there are many possibilities represented by the grey sphere). From this, it can be understood that the descent method tries to converge towards x. It perhaps reaches an inner sphere, Ψ, that represents MRI data corruption which cannot be removed. DDPMs, without guidance, would converge to any centroid that exists on the hyperplane. Guided DDPMs on the other hand are descending towards x but only ever perform a single step into the hypersphere of x itself. As a result, it can be interpreted as only ever being able to perform approximately a single descent step towards x. However, since the descent is a stochastic process (DDPM), it can reach a bigger range of different points within this hypersphere with each direction of approach pointing towards a noisy realisation of the output image i.e. it can approach any point of Ψ but it does not necessarily get as close to it as the traditional descent scheme. If the traditional descent scheme for MRI optimisation can be combined with stochastic gradient ascent for probabilistic modelling, then any point within Ψ itself can be reached (or at least much closer).

7.4.1 Score function decomposition for DDPM-based proximal reconstruction

In this section, we discuss an interpretation of DDPM-based accelerated MR image reconstruction and subsequently propose a novel, more direct approach to this task. As shown in section 2.4, the ELBO of DDPMs can be maximised by training with an L2 loss, equation (7.11), on the score function for each latent representation in the diffusion process.

$$||\epsilon_\theta(x_t, t) - \epsilon||^2, \tag{7.11}$$

where θ are the network parameters, ϵ_θ is the weighted score function learned by the network, x_t is the latent representation being scored, t is the current step in the (reverse) diffusion process and ϵ is the normally distributed noise that is used to formulate x_t (at training time). For reference, we repeat the formulation of x_t in equation (7.12) but advise the reader to refer to section 2.4 and [161] for more details on DDPMs.

$$x_t = \sqrt{\tilde{\alpha}}x + \sqrt{1 - \tilde{\alpha}}\epsilon, \tag{7.12}$$

where x is an image from the training set and $\tilde{\alpha}$ is related to the DDPM variance schedule, controlling the level of diffusion noise at each step of the diffusion process.

In order for the DDPM to generate images that are reconstructions of x_u, we condition the score function on x_u. The loss function for the DDPM then becomes equation (7.16).

$$||\epsilon_\theta(x_t, t, x_u) - \epsilon||^2 \tag{7.13}$$

In order to understand how the DDPM network, ϵ_θ, might use x_u in the scoring, we decompose the score function into two parts as shown in equation (7.14). In particular, we suggest that the DDPM network implicitly learns a hidden reconstruction of x_u which we denote as $\hat{x}_h = f'_{\theta'}(x_u)$, where f' is a hidden, internal function of the DDPM network, ϵ_θ, and θ' is a subset of θ.

$$\epsilon_\theta(x_t, t, x_u) = \frac{x_t - \sqrt{\tilde{\alpha}}\hat{x}_h}{\sqrt{1-\tilde{\alpha}}} + \epsilon'_\theta(x_t, t, \hat{x}_h), \tag{7.14}$$

where ϵ'_θ is another hidden, internal function of the DDPM network that does the actual learning of the reconstruction variability. In this decomposition, the first term represents a crude estimate of the diffusion noise based on trying to reconstruct x_u with a NN mapping, f', and the second term represents the probabilistic modelling of the reconstruction such as thermal noise, reconstruction possibilities and other variability present in the dataset.

By introducing $\hat{x}_h$ as a hidden auxiliary variable, it is possible to understand the DDPM loss (7.13) as minimising the loss between $\hat{x}_h$ and the image x with the DDPM modelling the data variability (and other uncertainty). Substituting (7.12) into equation (7.14) and then into the loss (7.13) produces equation (7.15) which highlights this understanding.

$$||\frac{\sqrt{\tilde{\alpha}}}{\sqrt{1-\tilde{\alpha}}}(\hat{x}_h - x) + \epsilon'_\theta(x_t, t, \hat{x}_h)||^2 \tag{7.15}$$

In other words, the DDPM is modelling the distribution of possible reconstructions after there is a hidden reconstruction $\hat{x}_h$ formed implicitly within the neural network, θ, (which in this case is a U-net).

7.4.1.1 Extension to multiple proximal steps: decomposed score function

Rather than conditioning the DDPM on simply x_u through a single (internal/implicit) reconstruction step, we can condition it on the output of an unrolled reconstruction network with N iterations, $\hat{x}_N^{\text{rec}} = f_\phi(x_u)$, where f is the reconstruction network and ϕ are its parameters. The loss function for the DDPM then becomes equation (7.16).

$$||\epsilon_\theta(x_t, t, \hat{x}_N^{\text{rec}}) - \epsilon||^2 \tag{7.16}$$

Model: Exp (Samples)	PSNR	SSIM
Supervised: DDPM	$\mathbf{35.55 \pm 1.62}$	$\mathbf{0.982 \pm 0.007}$
Supervised: Baseline	34.99 ± 1.49	0.977 ± 0.006
Unsupervised: DDPM (N_s=10)	$\mathbf{36.02 \pm 2.13}$	$\mathbf{0.980 \pm 0.010}$
Unsupervised: Baseline	33.01 ± 1.12	0.964 ± 0.008
Unsupervised: DDPM (N_s=1)	35.03 ± 1.93	0.968 ± 0.010

Table 7.2: Results of our study using a direct approach for performing MRI reconstruction with an iterative decomposition combined with the DDPM. The conditioning input to the DDPM is an N-cascade of Unets with learnable data consistency with $N = 5$. The baseline consisted of the same cascade but with $N = 6$. The supervised models are trained with fully sampled labels with a vanilla data preparation - undersampled acquisition as input, fully sampled label as output. The unsupervised models are trained using a data preparation similar to [177] due to the lack of fully sampled data, described in sections 2 and 3.1. In particular, it should be noted that the DDPM model only predicts a single k-space line at a time making inference significantly slower than the baseline.

Similar to the case without f, s_θ, may use the reconstruction $\hat{x}_N^{\text{rec}}$ in the scoring and subsequently, implicitly optimise the network f_ϕ. Similar to (7.14), we decompose the score function into two parts as shown in equation (7.17).

$$\epsilon_\theta(x_t, t, \hat{x}_N^{\text{rec}}) = \frac{x_t - \sqrt{\tilde{\alpha}}\hat{x}_N^{\text{rec}}}{\sqrt{1-\tilde{\alpha}}}\epsilon'_\theta(x_t, t, \hat{x}_N^{\text{rec}}) \tag{7.17}$$

In this decomposition, optimising the score function would optimise the unrolled network f in such a way that benefits the probabilistic modelling of ϵ' and the overall scoring function ϵ. This is a direct way to harness exist unrolled network architectures with DDPMs.

We had to use a gradient clipping value of 0.5 for the gradient norm due to the $\sqrt{1-\tilde{\alpha}}$ term in the denominator of equation (7.15) causing large, stochastic gradient spikes that led to unstable training. U-nets are used for each cascade in the initial reconstruction model and they are also used for the DDPM. Each U-net consists of 8.6M parameters each. The number of parameters for our proposed method and control experiments are the same.

7.4.2 Experimental results

Using an unrolled DC-CNN reconstruction network appended to the DDPM to form a conditional DDPM model, we formulate a cascading DDPM. It should be noted that DC-CNN is only one class of possible unrolled iterative networks that can be used — we choose DC-CNN here for its simplicity but in principle any type of decomposition can be used such as a gradient descent approach as in variational networks [110].

In our experiments, we use $N_c = 5$ cascades for the DC-CNN network. The output of this network is used to condition the DDPM which consists of the same U-net from section 7.3. Each cascade of the DC-CNN also uses the same U-net except without the embedding layer used for conditioning with the diffusion index, t. In the case of the baseline network, we use $N_c = 6$ cascades for a fair comparison since the DDPM represents a proximal mapping/reconstruction step.

The cascading DDPM is tested in both a supervised and unsupervised scenario with the results shown in table 7.2. In the supervised case, we perform obtain a single reconstruction from the DDPM for each example in the test set ($N_{\text{it}} = 1$) and use this in our comparisons against the ground truth. In the unsupervised case, we show the quantitative results for a single reconstruction from the DDPM as well as for $N_{\text{it}} = 10$ reconstructions from the DDPM which are then subsequently averaged.

The cascading DDPM shows its benefit over the ablation baseline for an acceleration rate of $\times 4$ in both the supervised and unsupervised scenario however more aggressive rates should be investigated (see section 7.6 for more studies). Figures 7.4-7.11 show example reconstructions of the cascading DDPM.

7.4.3 Discussion

The cascading DDPM approach to creating a higher fidelity MRI reconstructor has been shown to work quantitatively however the cost in forward inference is much more expensive. The

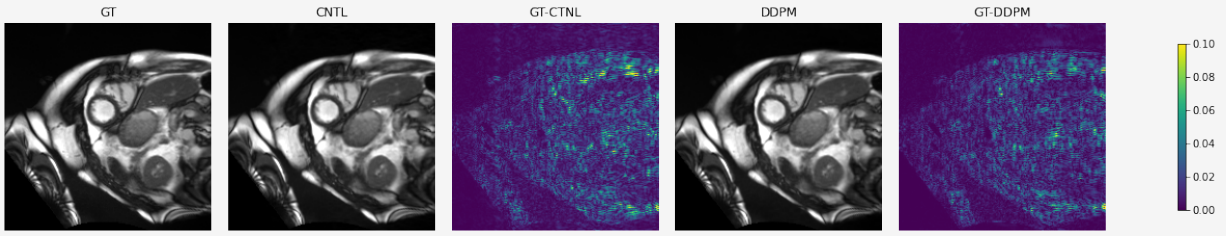

Figure 7.4: Example reconstruction from the UK BioBank dataset in the supervised setting. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. CNTL - PSNR: 34.84, SSIM: 0.979. DDPM - PSNR: 35.84, SSIM: 0.983.

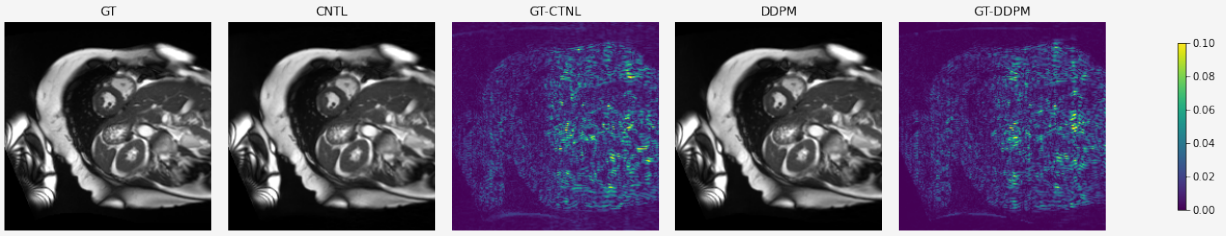

Figure 7.5: Example reconstruction from the UK BioBank dataset in the supervised setting. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. CNTL - PSNR: 34.79, SSIM: 0.978. DDPM - PSNR: 35.38, SSIM: 0.982.

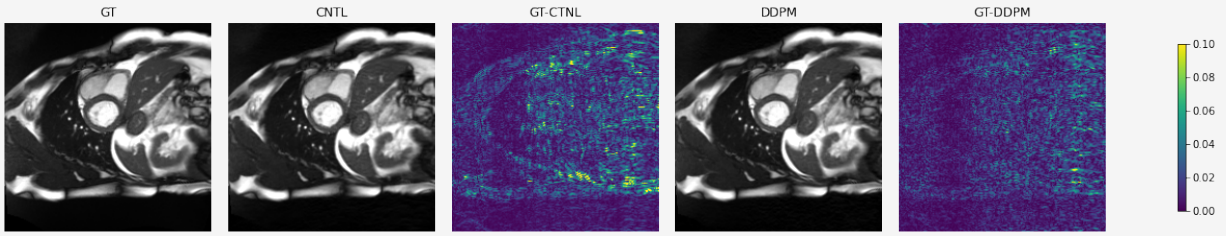

Figure 7.6: Example reconstruction from the UK BioBank dataset in the unsupervised setting. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. Only a single candidate reconstruction from the DDPM is used for comparison, $N_s = 1$. CNTL - PSNR: 32.76, SSIM: 0.969. DDPM - PSNR: 34.98, SSIM: 0.976.

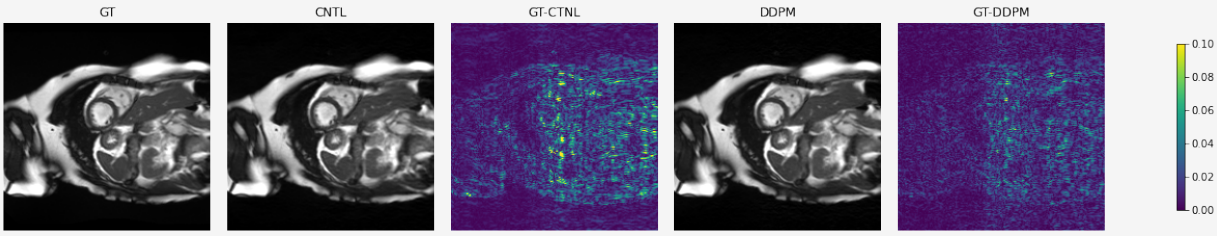

Figure 7.7: Example reconstruction from the UK BioBank dataset in the unsupervised setting. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. Only a single candidate reconstruction from the DDPM is used for comparison, $N_s = 1$. CNTL - PSNR: 33.51, SSIM: 0.970. DDPM - PSNR: 35.93, SSIM: 0.976.

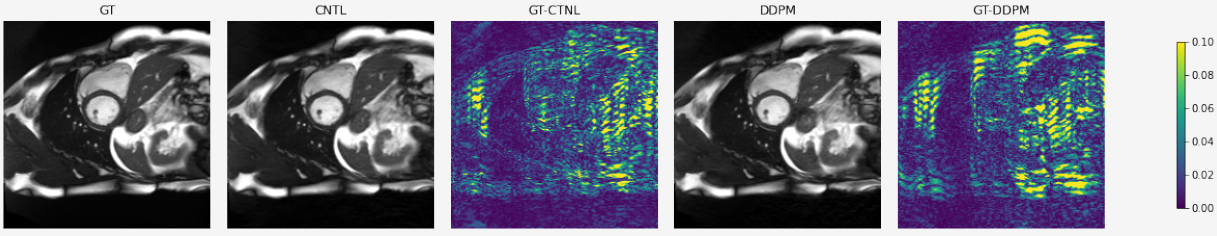

Figure 7.8: Example reconstruction from the UK BioBank dataset in the unsupervised setting. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. Only a single candidate reconstruction from the DDPM is used for comparison, $N_s = 1$. In this example, the DDPM produces a worse reconstruction compared with the CNTL. CNTL - PSNR: 29.46, SSIM: 0.943. DDPM - PSNR: 28.10, SSIM: 0.933.

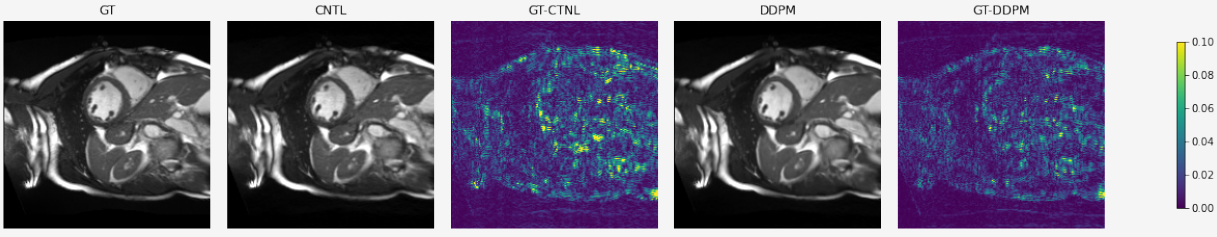

Figure 7.9: Example reconstruction from the UK BioBank dataset in the unsupervised setting. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. Multiple candidate reconstruction from the DDPM are sampled and then averaged to produce a single reconstruction, $N_s = 10$. CNTL - PSNR: 32.03, SSIM: 0.954. DDPM - PSNR: 34.46, SSIM: 0.972.

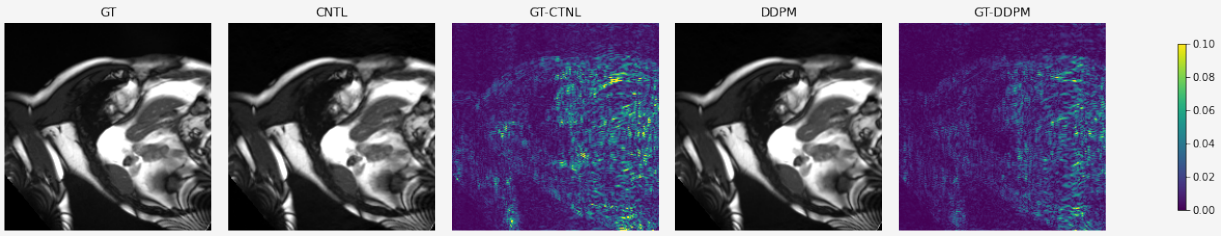

Figure 7.10: Example reconstruction from the UK BioBank dataset in the unsupervised setting. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. Multiple candidate reconstruction from the DDPM are sampled and then averaged to produce a single reconstruction, $N_s = 10$. CNTL - PSNR: 32.93, SSIM: 0.973. DDPM - PSNR: 35.41, SSIM: 0.984.

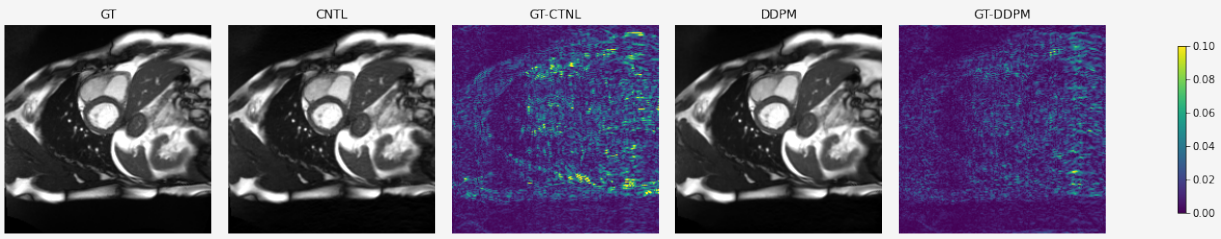

Figure 7.11: Example reconstruction from the UK BioBank dataset in the unsupervised setting. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. Multiple candidate reconstruction from the DDPM are sampled and then averaged to produce a single reconstruction, $N_s = 10$. CNTL - PSNR: 32.76, SSIM: 0.969. DDPM — PSNR: 36.16, SSIM: 0.984.

forward inference of the baseline is a matter of milliseconds compared with the 30 seconds of the cascading DDPM using an NVIDIA RTX 2080. This is due to the fact that the network must be evaluated $T = 1000$ times before generating a candidate reconstruction. However, it should be noted that reducing inference time is an active area of research with respect to DDPMs. For example, [181] focuses on a contraction theory approach to show that a one-step forward diffusion of an initial estimate can be used to significantly decrease the number of required reverse diffusion steps [181, 43, 11]. We leave this possible research direction for future work.

7.5 Handling and mitigating for corrupted data

In this section, the use of DDPMs to handle corruption acquisitions is investigated with fully-sampled targets. This serves as a preliminary proof of concept for the scenario without fully-sampled targets in section 7.6.

In the scenario where ground truth data does not exist due to the presence of noise in the imaging targets, training a model to produce high fidelity reconstructions becomes increasingly cumbersome. As mentioned in section 7.1, vanilla loss functions may not be appropriate training objectives for noisy or corrupted targets. Instead, we propose off-loading the corruption to a probabilistic model.

Generative models have been extensively studied in the deep learning and are still an active area of research. The most popular flavours of generative models are GANs and WGANs. However, it has been shown that whilst these models can produce high quality outputs (e.g. StyleGAN [128]), the curse of dimensionality has masked their ability to learn simple distributions. [99] studies the ability of GANs and WGANs to learn simple 1D parametric distributions. They find that these flavour of GANs fail to reproduce the target distribution when sampling at test time. They can generate samples around modal points of the distribution but fail to capture its true properties. They do find a particular flavour of GAN called Maximum Mean Discrepancy (MMD) GAN to learn the studied parametric distributions well [88]. However, the outputs

of this model are not of high fidelity and would be inappropriate for our task where it is of the utmost importance that the reconstructions remains faithful to the acquired data. [184] shows that high fidelity images can be generated from DDPMs and are straight forward to train alleviating the issues of mode-collapse and discriminator network tuning that GANs experience.

We propose using a DDPM as a generative model that learns the underlying distribution of the data corruption in an accelerated acquisition setting. In this section, we focus on the scenario of x4 undersampled acquisitions and fully-sampled imaging targets which both contain non-Gaussian noise for the network to learn. This scenario is chosen as a sanity check for the use of DDPMs to deal with corrupted data with non-Gaussian noise in the imaging target e.g. motion corrupted data.

The choice of non-Gaussian distribution in this study is the Rician distribution. We choose the Rician distribution since it is simple and has some known unbiased estimators [13]. Furthermore, it also draws parallels to thermal noise that is typical in MRI scanners but only if the imaging targets were magnitude images such as if training using DICOM images. With DICOM images, raw data is discarded from the scanner such as with the DICOM data that form part of the fastMRI challenge [120]. Magnitude images as targets have also been used in deep learning networks such as in AUTOMAP [121]. One could also conceive a network which is fed multi-coil data and subsequently designed to directly reconstruct the multi-coil RSS image.

Furthermore, we do not keep the noise level in the target images constant. One of the assumptions of using an L2 to perform maximum likelihood estimation is that we have identically, independently distributed normal errors. Not only have we changed the error from a normal distribution to Rician, but we also vary the pixel-wise noise level between 0.001 to 0.01 with the upper bound being large enough to generate surprisingly noisy-looking target images as shown in Figure 7.12.

In this study, we use the magnitude images, M, from the UK BioBank study. The k-space data is simulated using a synthetic phase with an added Gaussian noise to both the real and imaginary components [116]. The magnitude image of this k-space data, $\tilde{M}$, would subsequently

have Rician-distributed noise as shown in Equation (7.18):

$$\tilde{M} \sim \text{Rician}\left(M, \sigma\right) \tag{7.18}$$

where σ is the noise level and M is the magnitude image. Equation (7.19) shows how $\tilde{M}$ is formed from Gaussian noise in the real and imaginary components of the complex image:

$$\tilde{M} = \sqrt{X^2 + Y^2}, \tag{7.19}$$

where

$$X \sim N\left(M\cos\theta, \sigma^2\right), Y \sim N\left(M\sin\theta, \sigma^2\right), \tag{7.20}$$

and θ is the phase of the image.

For the purposes of this preliminary investigation, the imaging targets for the network to learn is the non-Gaussian noisy magnitude image $\tilde{M}$. Whilst in practice we could train using the complex valued images with Gaussian noise, the purpose of this investigation is to demonstrate the ability of DDPMs to better mitigate for non-Gaussian noise.

One possible method of reconstructing an image with noise is to perform maximum likelihood estimation of the parameters of the underlying distribution. This would require sampling many noisy reconstructions from the DDPM and using them in an appropriate estimator.

In our case, we know that this distribution is Rician and hence can estimate the parameter of interest, the image, from an estimator of the second moment which is unbiased [13]:

$$\hat{x}_c = \sqrt{\langle f_{\text{ddpm}}(y)^2\rangle - 2\sigma^2}, \tag{7.21}$$

where f_{ddpm} is the output from the DDPM model and σ is an estimate of the noise level which can be estimated empirically at inference since it isn't required in the training of the generative

model. (In terms of reconstruction for qualitative purposes such as visual inspection, knowledge of σ is not necessary and can be set to zero).

7.5.1 Method

A DDPM was trained with heavily noisy data with examples demonstrating the effective size of this noise in Figure 7.12. We use the magnitude images, M, from the UK BioBank study. The k-space data is simulated using a synthetic phase with an added Gaussian noise to both the real and imaginary components [116]. This results in a Rician noise in the imaging targets of the network. At training time, the DDPM never sees a noiseless image and hence in this sense, since we have no ground truth/noiseless labels, the task is unsupervised. At inference, the DDPM model is sampled several times ($N = 60$) under the same input condition, the undersampled, noisy acquisition. Using a-priori knowledge of the noise level, we use equation (7.21) to obtain an estimate of the noiseless image.

We used the cascading DDPM model as presented in section 7.4. Since the prior DC-CNN reconstruction network f use data consistency, we must prevent the noise parameter of DC-CNN approaching $\lambda \longrightarrow \infty$ which would likely prevent the DDPM from learn the corruption distribution at the acquired k-space points and thus would experience unconventional mode collapse. We investigate two strategies to counter this mode collapse:

- **Scenario 1: Double Corruption** - We have acquired fully-sampled but noisy targets. The training inputs are retrospectively undersampled images of the noisy fully-sampled acquisitions which have an additional Gaussian noise manually added pixel wise so that at training time, the learned denoising step does not force the data consistency to $\alpha = \infty$ (mode collapse). (See equation 7.3 for more details on data consistency). In the case of other data corruptions, such as motion corruption, other strategies must be investigated to ensure that we reasonably maximise the entropy of the total noise in the input images given the noisy output.

- **Scenario 2: Double Acquisition** - This method requires two acquisitions of our data.

We have acquired fully-sampled but noisy targets acquired twice sequentially. The training inputs are retrospectively undersampled images of one of the noisy targets and then in the target used in training is the other noisy acquisition. We hypothesise that this method would perform better than in scenario 1 since we do not need to further corrupt our training with noise. The case of $\alpha = \infty$, i.e. mode collapse, is automatically avoided since the noise component of the values between the two acquisitions at each k-space point are different but drawn from the same noise distribution (Rician in this case). In the case of other data corruptions, such as motion corruption, this would simply require collecting the data twice.

The baseline network is the DC-CNN reconstruction network with an extra cascade. The magnitude image from the complex output of the DC-CNN is formed and trained with an L2 loss against the fully-sampled noisy magnitude image, $\tilde{M}$.[2]

7.5.1.1 Faster sampling

Properties of the Rician distribution It should be noted that in areas of high SNR, the Rician distribution is close to Gaussian. The mean, and thus modal point of this distribution is $\sqrt{y^2 + \sigma^2}$. In areas of low SNR (e.g. zero), the distribution is Rayleigh distributed with a mode of σ. If we have easy access to the statistical properties of the data-corruption distribution learnt by the network, using the mode of this distribution, we can make estimates for y.

Extraction of modal point from DDPMs The DDPM distribution learned should be mostly representative of the noise distribution in the data. Thus, the DDPM should be approximately Rician distributed. At inference time, the output of the DDPM is generated through the equation (2.21). This iterative inference uses the current latent point, calculates the score function to move to an area of higher likelihood and performs a random (normally distributed) step. We hypothesise that we can obtain an estimate of the mode of the DDPM by using the modal points at each step of reverse diffusion process. The modal point of the normally

[2]It should be noted that conventionally, the loss would be calculated on the complex output since the noise in the target would Gaussian and the network optimisation is more convex. However, the choice of training objective here is justified due to the purpose of this preliminary investigation.

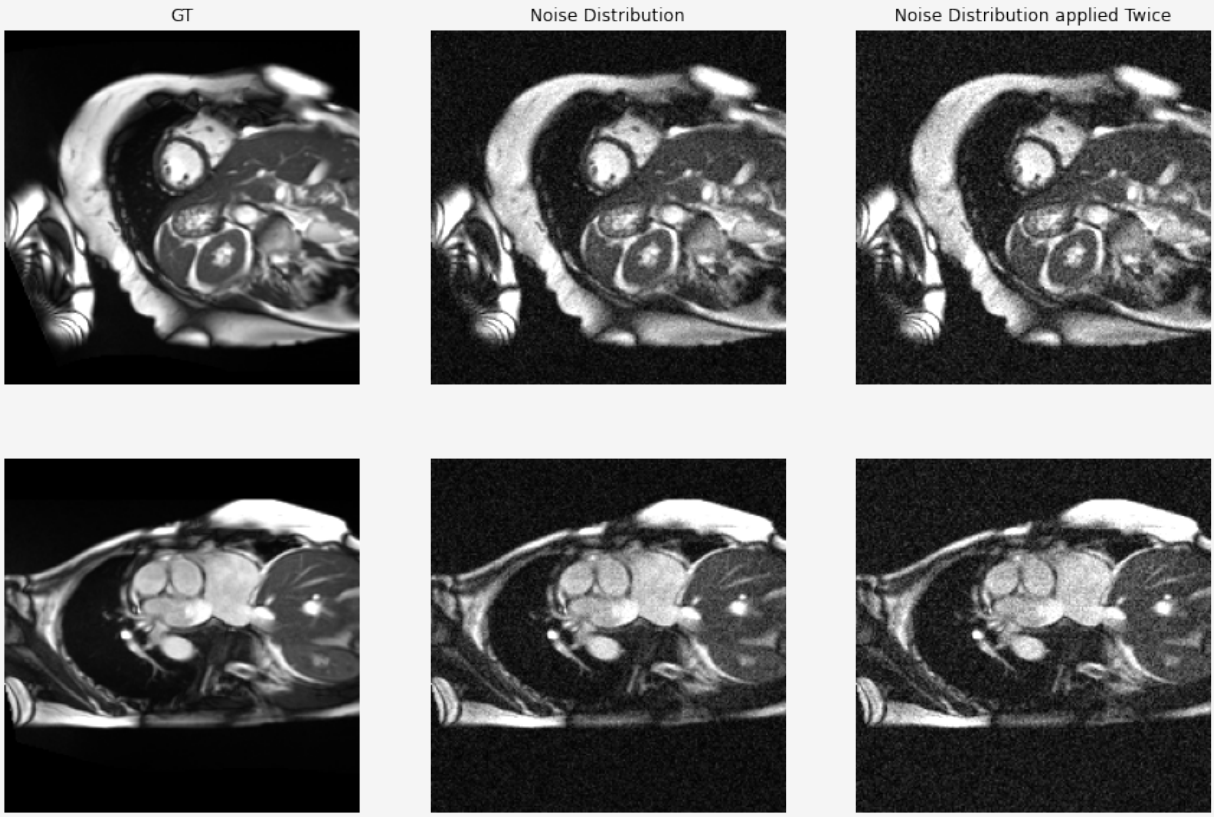

Figure 7.12: Examples of the maximum amount of noise applied to images from the UK BioBank. On the left is the ground truth magnitude image, middle shows the image with the noise distribution applied, right shows the image with the noise distribution applied twice (see Scenario 1).

distributed walk is simply zero. Thus by removing the random walk, we only need a single sample of the DDPM (with zero walk) to estimate the mode of the distribution.

7.5.2 Results

We investigate the ability of the DDPM to reconstruct noiseless images under two methods of sampling from the DDPM:

- $N_s = 60$ - We take 60 samples from the DDPM and use equation (7.21) to generate a noiseless reconstruction
- $N_s = 1$ - We take 1 sample from the DDPM but set the volatility term in the reverse diffusion SDE to zero in an effort to sample close to the modal point of the corruption distribution.

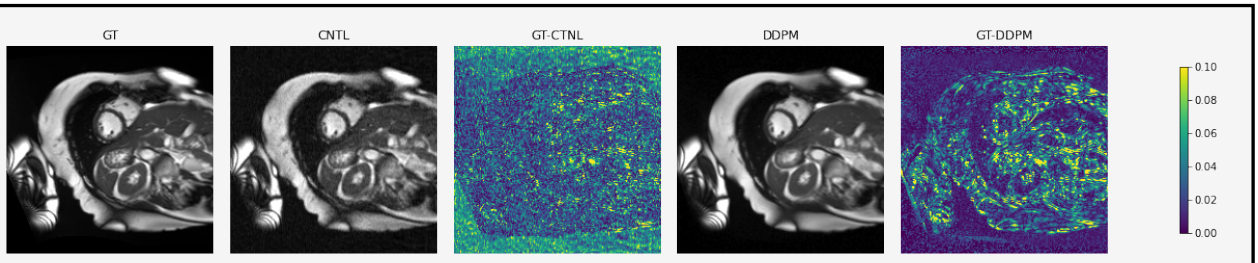

Figure 7.13: Example reconstruction from the UK BioBank dataset with scenario 2 setup. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. $N_s = 60$ candidate reconstructions from the DDPM are sampled and then equation (7.21) is used to generate a single reconstruction. CNTL - PSNR: 27.16, SSIM: 0.770. DDPM - PSNR: 28.92, SSIM: 0.912.

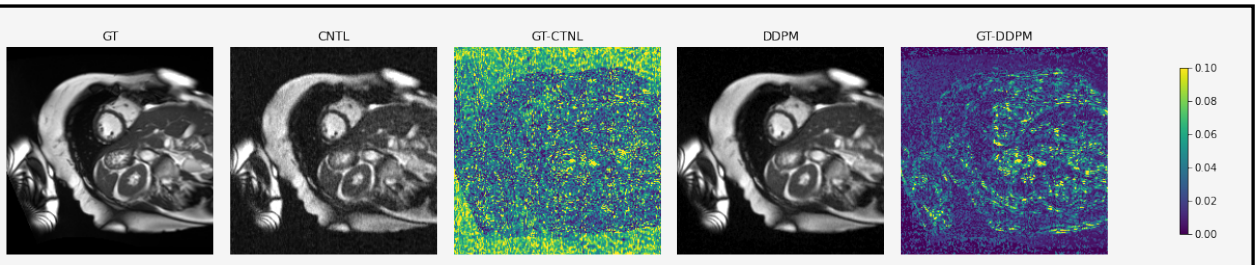

Figure 7.14: Example reconstruction from the UK BioBank dataset with scenario 1 setup. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. $N_s = 60$ candidate reconstructions from the DDPM are sampled and then equation (7.21) is used to generate a single reconstruction. CNTL - PSNR: 24.64, SSIM: 0.704. DDPM - PSNR: 29.97, SSIM: 0.915.

This is described above in section 7.5.1.1.

Furthermore, the DDPM generates these reconstructions with acquisitions that contain maximum noise - examples of the inputs to the DDPM as shown in Figure 7.12. Table 7.3 and Figures 7.13-7.20 shows evaluations of the cascading DDPM in the scenario of non-Gaussian noise heavily corrupting the acquisition and target data.

7.5.3 Discussion

We show that noiseless reconstructions can be obtained with high qualitative and quantitative performance. These noiseless estimates were facilitated by using a-priori knowledge of the type of noise distribution which was Rician. In practice, a-priori knowledge of the noise level

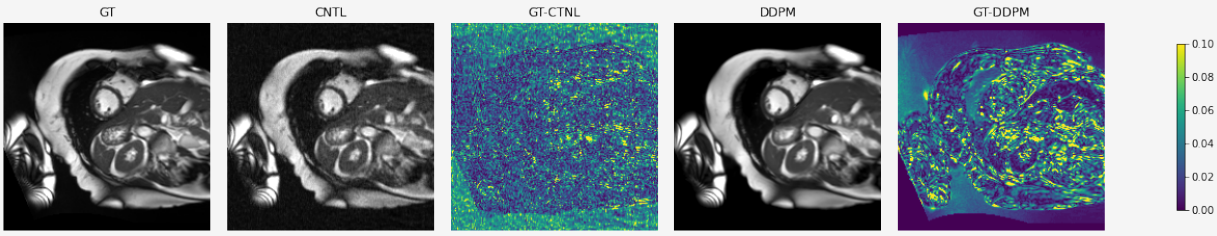

Figure 7.15: Example reconstruction from the UK BioBank dataset with scenario 1 setup. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. A single $N_s = 1$ candidate reconstruction from the DDPM is sampled using the methodology from section 7.5.1.1. Subsequently, equation (7.21) is used to generate a single reconstruction. CNTL - PSNR: 24.64, SSIM: 0.704. DDPM - PSNR: 28.45, SSIM: 0.878.

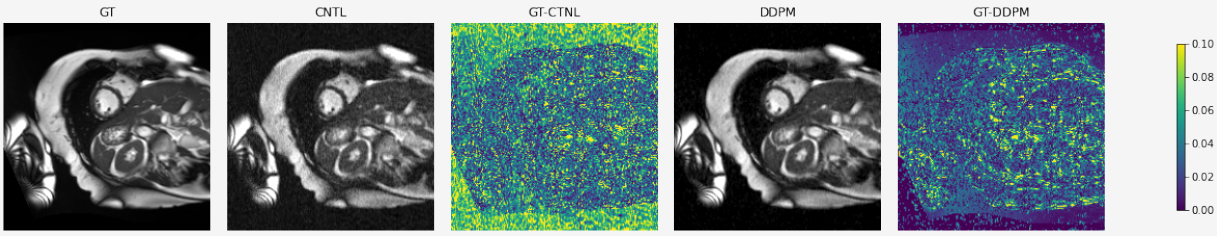

Figure 7.16: Example reconstruction from the UK BioBank dataset with scenario 2 setup. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. A single $N_s = 1$ candidate reconstruction from the DDPM is sampled using the methodology from section 7.5.1.1. Subsequently, equation (7.21) is used to generate a single reconstruction. CNTL - PSNR: 27.16, SSIM: 0.770. DDPM - PSNR: 27.81, SSIM: 0.842.

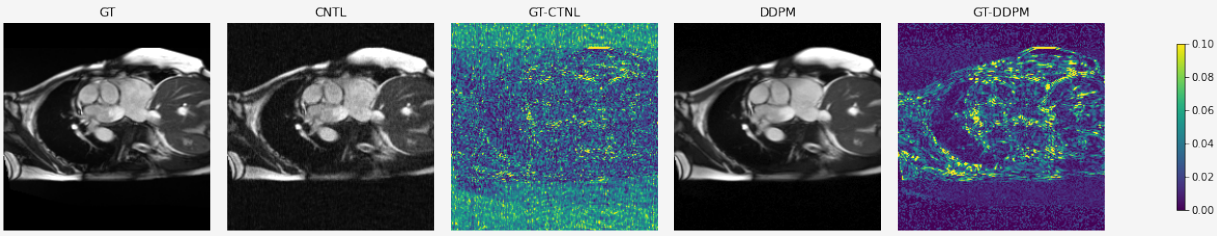

Figure 7.17: Example reconstruction from the UK BioBank dataset with scenario 2 setup. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. $N_s = 60$ candidate reconstructions from the DDPM are sampled and then equation (7.21) is used to generate a single reconstruction. CNTL - PSNR: 26.75, SSIM: 0.658. DDPM - PSNR: 29.58, SSIM: 0.907.

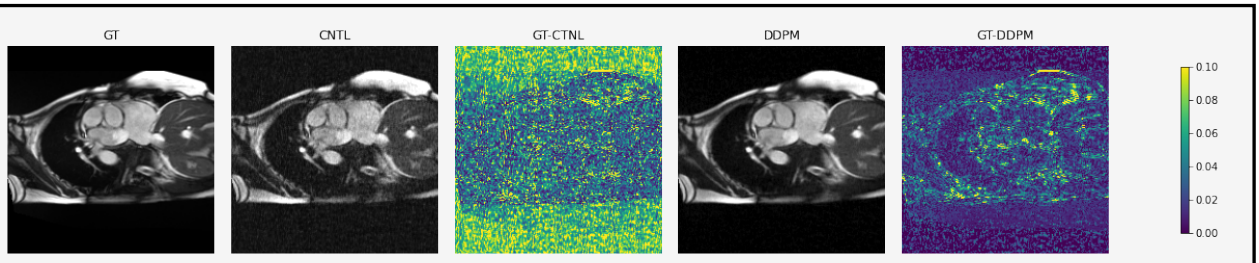

Figure 7.18: Example reconstruction from the UK BioBank dataset with scenario 1 setup. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. $N_s = 60$ candidate reconstructions from the DDPM are sampled and then equation (7.21) is used to generate a single reconstruction. CNTL - PSNR: 24.09, SSIM: 0.587. DDPM - PSNR: 30.48, SSIM: 0.893.

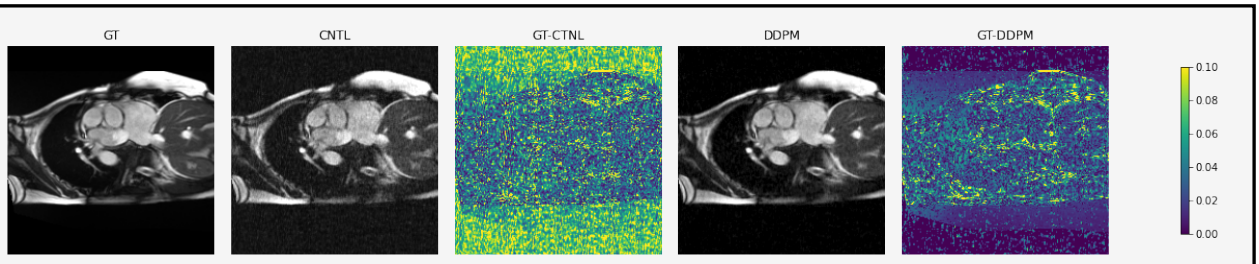

Figure 7.19: Example reconstruction from the UK BioBank dataset with scenario 1 setup. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. A single $N_s = 1$ candidate reconstruction from the DDPM is sampled using the methodology from section 7.5.1.1. Subsequently, equation (7.21) is used to generate a single reconstruction. CNTL - PSNR: 24.09, SSIM: 0.587. DDPM - PSNR: 29.09, SSIM: 0.872.

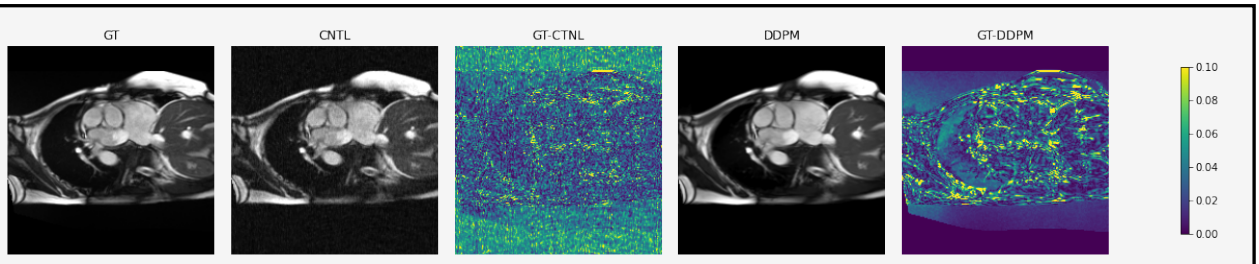

Figure 7.20: Example reconstruction from the UK BioBank dataset with scenario 2 setup. The CNTL is the DCCNN with $N_c = 6$ and the DDPM uses $N_c = 5$ followed by a DDPM. A single $N_s = 1$ candidate reconstruction from the DDPM is sampled using the methodology from section 7.5.1.1. Subsequently, equation (7.21) is used to generate a single reconstruction. CNTL - PSNR: 26.81, SSIM: 0.657. DDPM - PSNR: 28.62, SSIM: 0.862.

Model: Data (Samples)	PSNR	SSIM
DDPM: Double Corruption (N=60)	$\mathbf{29.96 \pm 0.80}$	$\mathbf{0.907 \pm 0.014}$
DDPM: Double Corruption (N=1)	28.32 ± 0.69	0.868 ± 0.014
Baseline: Double Corruption	24.57 ± 0.42	0.678 ± 0.065
DDPM: Double Acquisition (N=60)	$\mathbf{29.04 \pm 0.90}$	$\mathbf{0.907 \pm 0.014}$
DDPM: Double Acquisition (N=1)	28.13 ± 0.90	0.846 ± 0.021
Baseline: Double Acquisition	27.12 ± 0.44	0.750 ± 0.062

Table 7.3: Results when training with fully sampled acquisitions as targets but with a high Rician noise in the targets and inputs. Here the DDPM models use $N = 5$ cascades and the baseline models use $N = 6$ cascades. It can be seen here that the baseline performs poorly against the DDPM-based models due to an inappropriate noise distribution (Rician) in the target.

is difficult to obtain but with DDPMs, we can estimate it empirically by viewing the image background of the second-moment estimator.

We hypothesised that scenario 1, 'Double Corruption', would perform worse than in scenario 2, 'Double Acquisition'. In the case of the baseline, this is true for the aforementioned reasons. However, for our proposed method, scenario 1 and 2 perform comparably which indicates that our proposal for avoiding mode collapse is sufficient. It should be noted for unsupervised (not fully sampled) MRI reconstruction, these scenarios are not required since in the unsupervised DDPM, the target is not contained within the condition.

We note that formulation of an unbiased estimator for the noiseless image is only possible in this study as we have a-priori knowledge of the noise present in the data. However, many other types of noise are present in MRI data such as motion corruption. We presented the use of deterministic sampling - using a single sample - to extract a modal point of the data distribution that appears to mitigate for the data corruption without necessarily requiring the use of an unbiased estimator (since the aim is to obtain an image for clinical use). We note that the concept of deterministic sampling from DDPMs has been studied prior to this work in the form of denoising diffusion implicit models (DDIMs; [168]).

7.6 Application to fastMRI dataset

In this section, we combine the ideas and conclusions from the prior sections to consider the case of accelerated Cartesian MRI reconstruction without fully-sampled training data in the presence of thermal noise with real knee data from the fastMRI challenge, in both the single and multi-coil scenario. Specifically, we focus on self-supervised data consistent decomposed cascade DDPMs for the fastMRI dataset. The fastMRI dataset consists of more than 1500 knee images from a variety of different scanners at 1.5T and 3T field strength using 15 coils and a turbo-spin echo acquisition protocol. Information about the data can be found in [120].

The fastMRI dataset already contains an underlying non-identically distributed thermal noise that our approach would mitigate for. The fastMRI dataset contains fully sampled k-space hence we retrospectively generate unique undersampling masks for each subject in the dataset mimicking the scenario whereby fully-sampled k-space data may not exist. The DDPM model proposed never sees a fully-sampled image and yet one can generate one by estimating the k-space line by line (such as in section 7.3).

Whilst the fastMRI is inherently a multi-coil acquisition, single coil acquisitions are simulated using a linear combination of the coil acquisitions in a method from [175]. The fastMRI dataset already includes this simulated single-coil signal.

The fastMRI dataset is well known for containing a noisy background where the RSS combination of coils has resulted in a near chi-squared distribution of noise (or Rayleigh in the single coil case) [126]. The noise is also perceptible in the regions of interest in certain images in the dataset. However, in other images, the noise is less visible. There may also be less obvious noise in the form of motion corruption or scanner artefacts. Traditional losses may struggle to handle the nature of this distribution of reasons mentioned in section 7.1.

The DDPM approach outlined in section 7.5 provides a convenient method of handling such noisy data in order to obtain estimates of the true, noiseless image. In this aforementioned section, the targets were fully sampled which meant specific data acquisition scenarios were

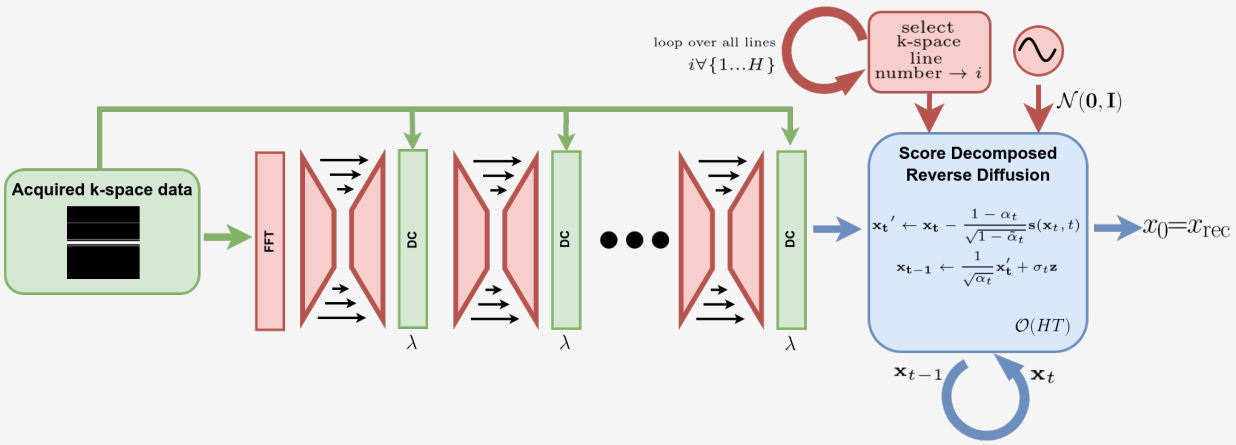

Figure 7.21: Illustration of the DC2DDPM methodology which makes use of a decomposed score function.

devised to validate the use of DDPMs in the case of noisy data (section 7.5). This study takes place in the unsupervised setting whereby fully-sampled targets do not exist. In section 7.3, at training time, the target k-space line is exclusive of the input data to the DDPM model. As a result, the 1-DDPM (with cascades) is naturally suited for use with noisy data in the partially-sampled data setting.

It should be noted that in this study, the target may indeed have a normal noise but is not necessarily identically distributed from one subject to the next. This provides another motivation for using DDPMs for MRI reconstruction.

7.6.1 Experimental method

We use the same U-net as in previous sections with the explicit score decomposition from section 7.4. That is to say, the cascade model prior to the DDPM learns the deterministic part of the score function whilst the DDPM part learns the probabilistic intricacies. We refer to our proposal model as the data-consistent decomposed cascade DDPM (DC2DDPM). This model is illustrated in Figure 7.21.

We also made changes to accommodate for the multi-coil nature of part of the fastMRI data. In the case of the prior cascade model, The coils of the data were the convolutional channel

inputs and outputs as is the case with parallel coil networks (PCN) [140]. The DDPM model subsequent to this were provided all channel outputs of the cascade model. However, the DDPM model was only required to generate a single line of k-space for a single coil. The coil selection of the DDPM was provided by appending the cascade output of the required coil to the input to the DDPM.

It should also be noted that different images have varying amounts of thermal noise (section 7.5) as can be seen in Figure 7.22.

To perform forward inference with our method, we could extract multiple samples and use a suitable estimator for our image reconstruction. However, due to the dimensionality of the fastMRI images, we restrict the number of samples to a single inference step for each k-space line by setting the noise at each reverse diffusion step to zero. This is an identical sampling process to denoising diffusion implicit models (DDIMs) [168] and also has similarities to the method outlined in section 7.5.1.1. We empirically find that the DDIM sampling process leads to better quality reconstructions.

7.6.2 Results

We evaluated our models on the single coil and multi-coil fastMRI knee validation dataset. We used 1000 examples which were each retrospectively undersampled with a variable-density Gaussian-distributed mask. For the single coil case, we used an acceleration rate of x4 and for the multi-coil case, we used an acceleration rate of x8. Figures 7.23-7.30 show some examples of the outputs from the proposed DDPM with 5 cascades and a baseline with 6 cascading U-nets based on the model from [177][3]. The quantitative results are displayed in table 7.4.

[3]The loss for this baseline is on the complex image (Gaussian noise in target), not the magnitude image (Rician noise in target) unlike in section 7.5.

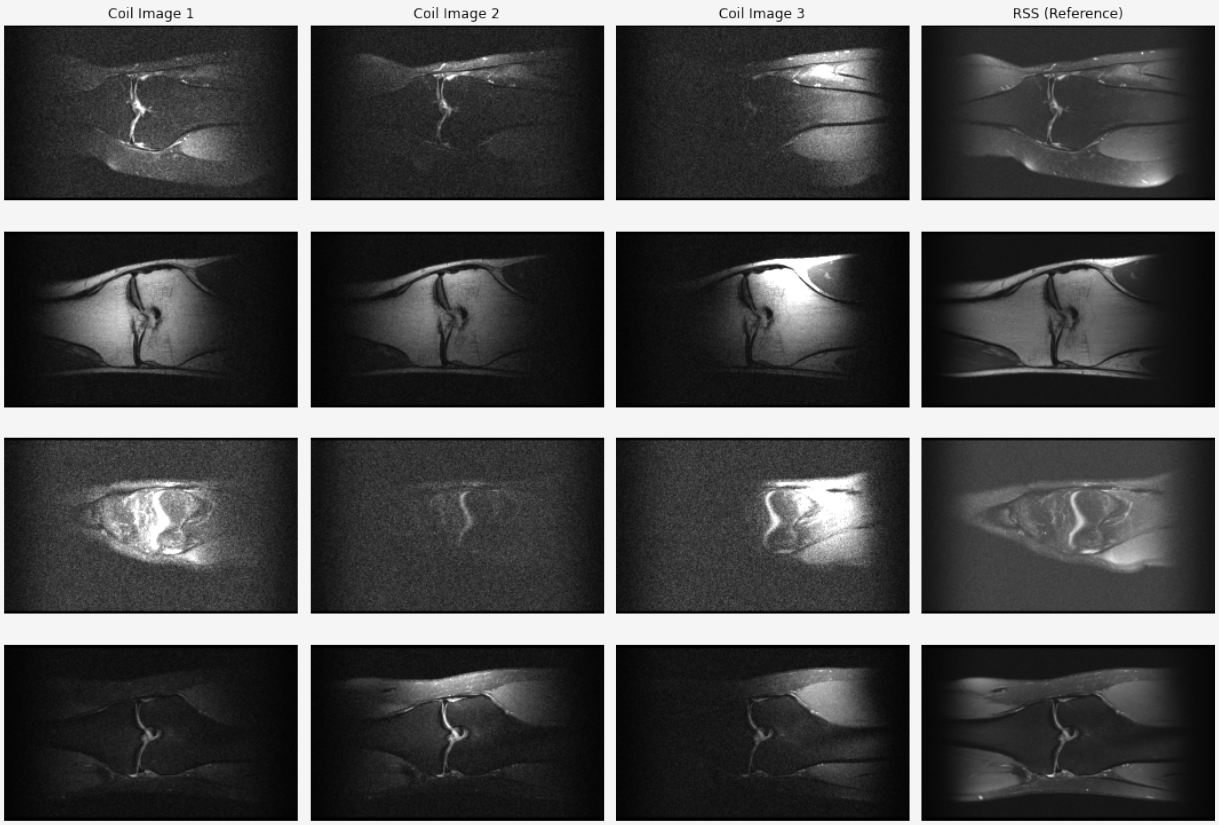

Figure 7.22: Examples of noisy images from the fastMRI dataset. In particular, note that different images have different levels of noise.

Experiment	**PSNR**	**ROI-PSNR**	**SSIM**	**ROI-SSIM**
CNTL (SC;x4) (Yaman et al. 2019)	$\mathbf{38.5 \pm 5.1}$	$\mathbf{37.9 \pm 4.8}$	0.833 ± 0.106	0.848 ± 0.096
1-DC2DDPM (SC;x4) [Ours]	38.4 ± 5.4	37.5 ± 5.2	$\mathbf{0.848 \pm 0.106}$	$\mathbf{0.850 \pm 0.103}$
CNTL (MC;x8) (Yaman et al. 2019)	28.8 ± 5.8	$\mathbf{29.1 \pm 5.8}$	0.715 ± 0.149	0.750 ± 0.135
1-DC2DDPM (MC;x8) [Ours]	$\mathbf{28.9 \pm 5.4}$	29.0 ± 5.3	$\mathbf{0.728 \pm 0.150}$	$\mathbf{0.751 \pm 0.134}$
Difference:				
1-DC2DDPM - CNTL (SC;x4)	-0.12 ± 0.40	-0.43 ± 0.49	-0.015 ± 0.021	$\mathbf{0.002 \pm 0.018}$
1-DC2DDPM - CNTL (MC;x8)	$\mathbf{0.14 \pm 0.50}$	-0.03 ± 0.56	$\mathbf{0.012 \pm 0.009}$	$\mathbf{0.002 \pm 0.007}$

Table 7.4: Table of results using 100 test images. The CNTL experiment refers to the model proposed in [177]. The PSNR and SSIM metrics are shown for the whole image as well as for a central crop of the image in the region of interest (ROI). The last two rows of the table show the average (and standard deviation) of the difference in PSNR and SSIM calculated per example in the test set. The multi-coil results are inconclusive - we were unable to reject the null hypothesis of greater 1-DC2DDPM performance over the CNTL with a Wilcoxon signed-rank test giving $p = 0.38$. The single-coil results suggest a quantitative advantage of the CNTL however we suggest that this is due to a lack of gold-standard data for evaluation as discussed in section 7.6.2.

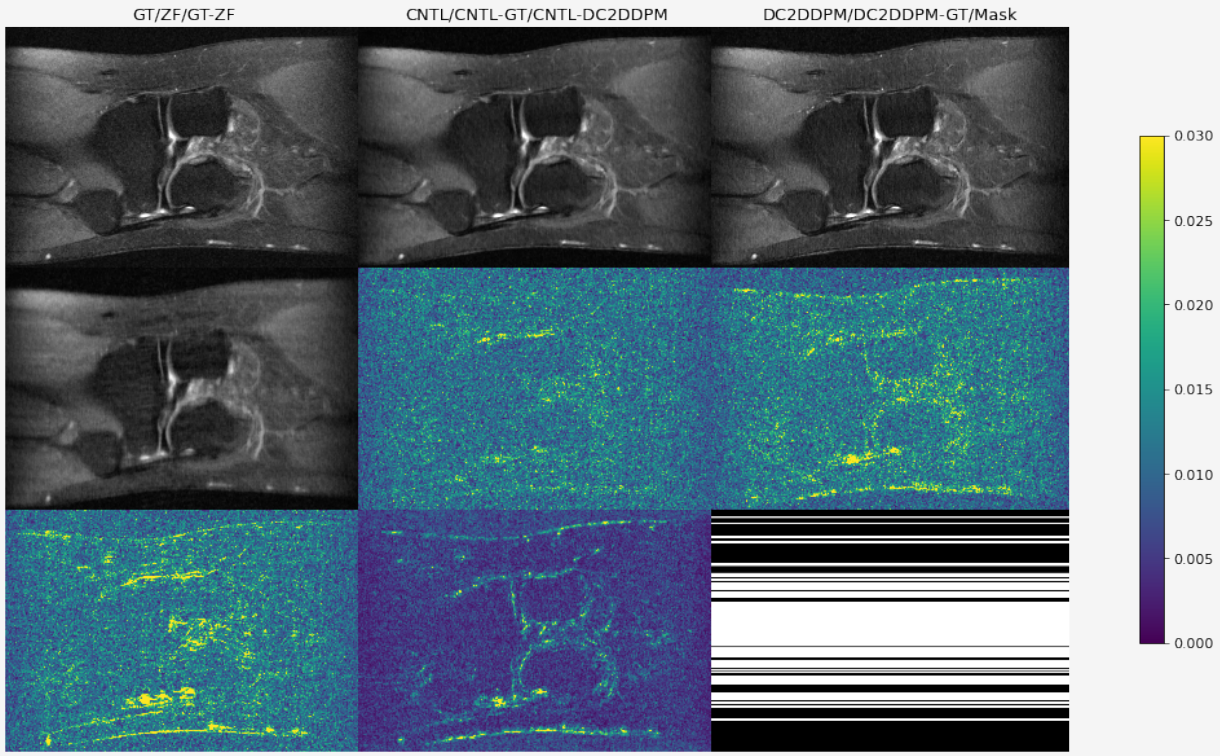

Figure 7.23: Reconstruction of x4 undersampled fastMRI single-coil knee image. (PSNR, SSIM) for CNTL and DC2DDPM: (36.8, 0.848), (36.0, 0.843).

7.6.3 Discussion

For the single coil data, the results shown in Figures 7.23-7.26 are subjective. Quantitatively from Table 7.4, it is clear that the DDPM does not outperform the baseline with a Wilcoxon signed-rank test giving $p = 0.49$ to reject the null hypothesis. This is a surprising result since qualitatively, the output reconstructions from the DDPM seem sharper and of a higher fidelity than both of the baseline and the ground truth reference. It is likely that the data corruption inherent in the fastMRI dataset itself leads to some image features becoming less sharp and increasingly blurred. The DDPM model tries to learn this corruption and when the modal sample is made from the DDPM, we actually recover the image without this corruption even with the DDPM having never seen an uncorrupted sample. This is the case that was hypothesised in section 7.5. Without better ground truth data, it is not possible to verify this and hence we suggest that for future work, a study with several radiologist opinions is conducted.

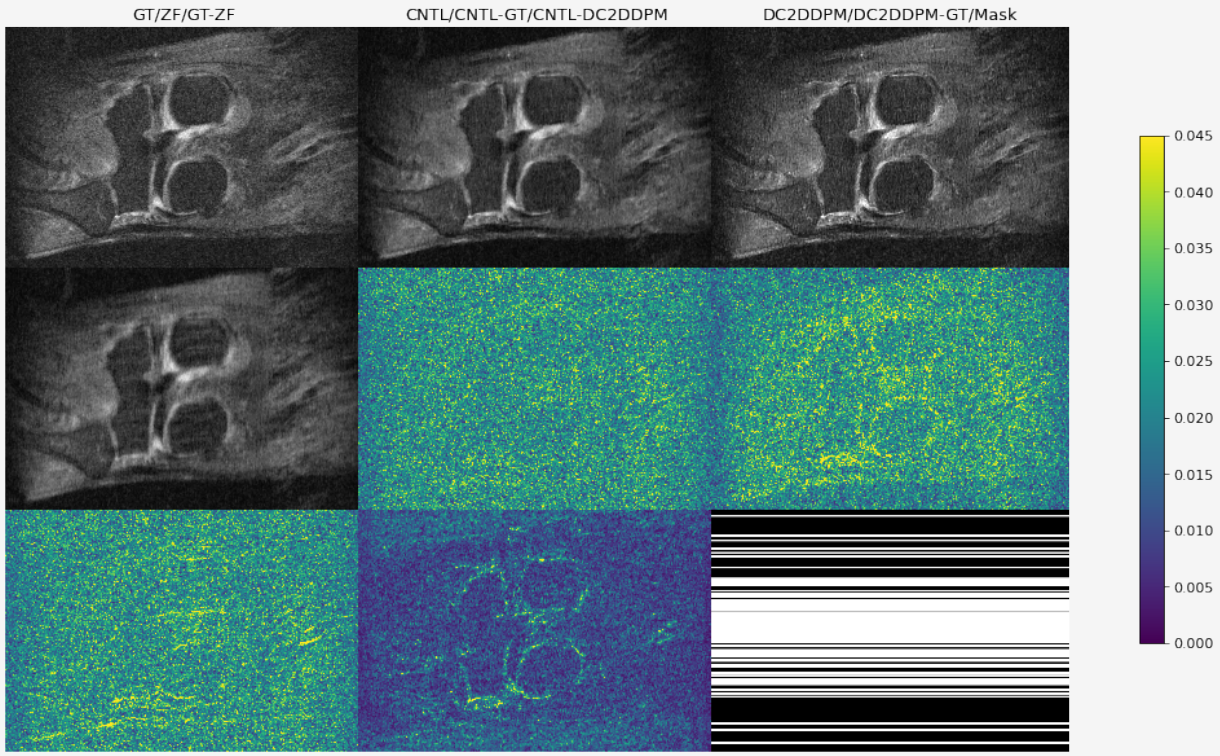

Figure 7.24: Reconstruction of x4 undersampled fastMRI single-coil knee image. (PSNR, SSIM) for CNTL and DC2DDPM: (30.9, 0.660), (30.4, 0.662).

The fact that the quantitative results do not reflect the sharper DDPM images highlights a number of problems in deep learning MRI reconstruction:

1. *Loss functions in deep learning* are currently somewhat misguided for the purpose of MRI reconstruction. In the case of supervised, discriminative training [19], the question of suitable loss functions becomes more complex with increasing image fidelity in the desired output. This is particularly the case where the target contains several, complex imperfections. A recent study using the fastMRI dataset explores the use of SSIM to optimise the output reconstruction [140]. In similar scenarios, it may be the case that probabilistic models become an increasingly more appropriate choice of reconstruction model.

2. *No gold-standard ground truth data* - The lack of this data makes it difficult to compare different reconstruction models. This was noted in the fastMRI challenge 2019 [126]. In reality, perfect ground truth data is never possible but the fastMRI data is particularly noisy compared to other (smaller) MRI image datasets currently available to researchers such as

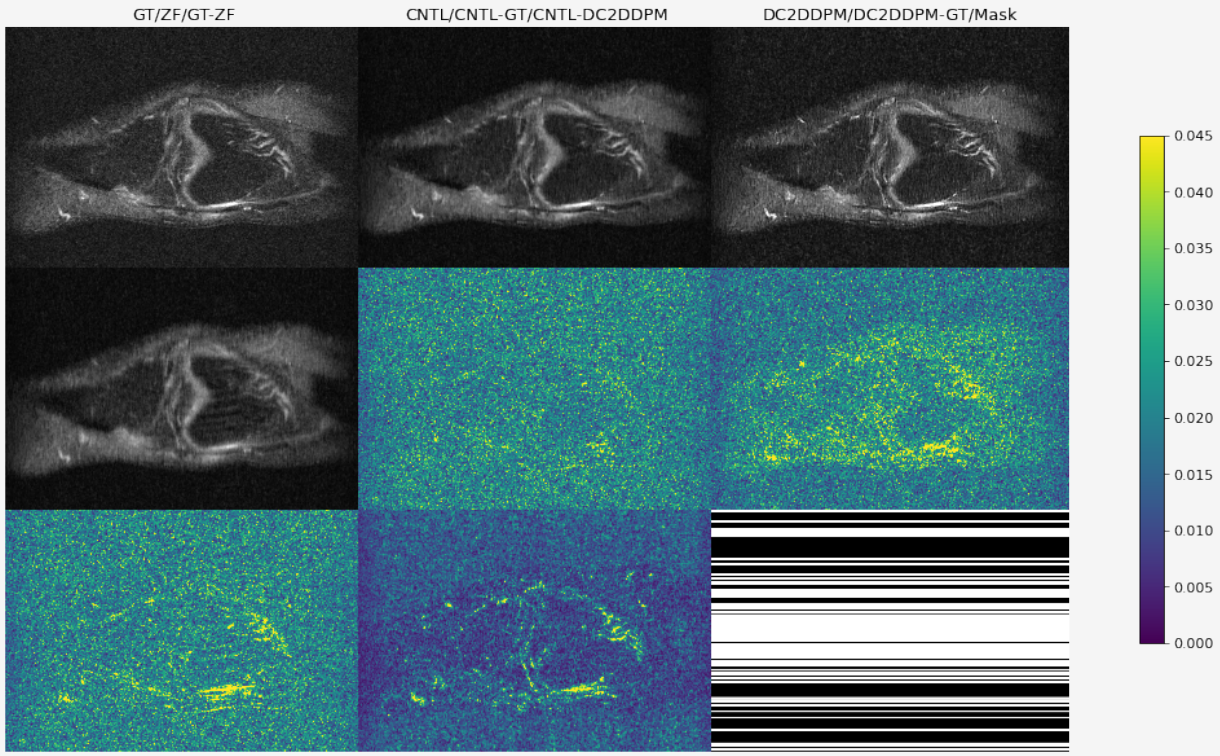

Figure 7.25: Reconstruction of x4 undersampled fastMRI single-coil knee image. (PSNR, SSIM) for CNTL and DC2DDPM: (31.7, 0.658), (31.6, 0.682). We note here that while the image features of the GT are more similar to that of the CNTL than in the DC2DDPM, the DC2DDPM reconstruction seems far sharper than that of both the CNTL and GT.

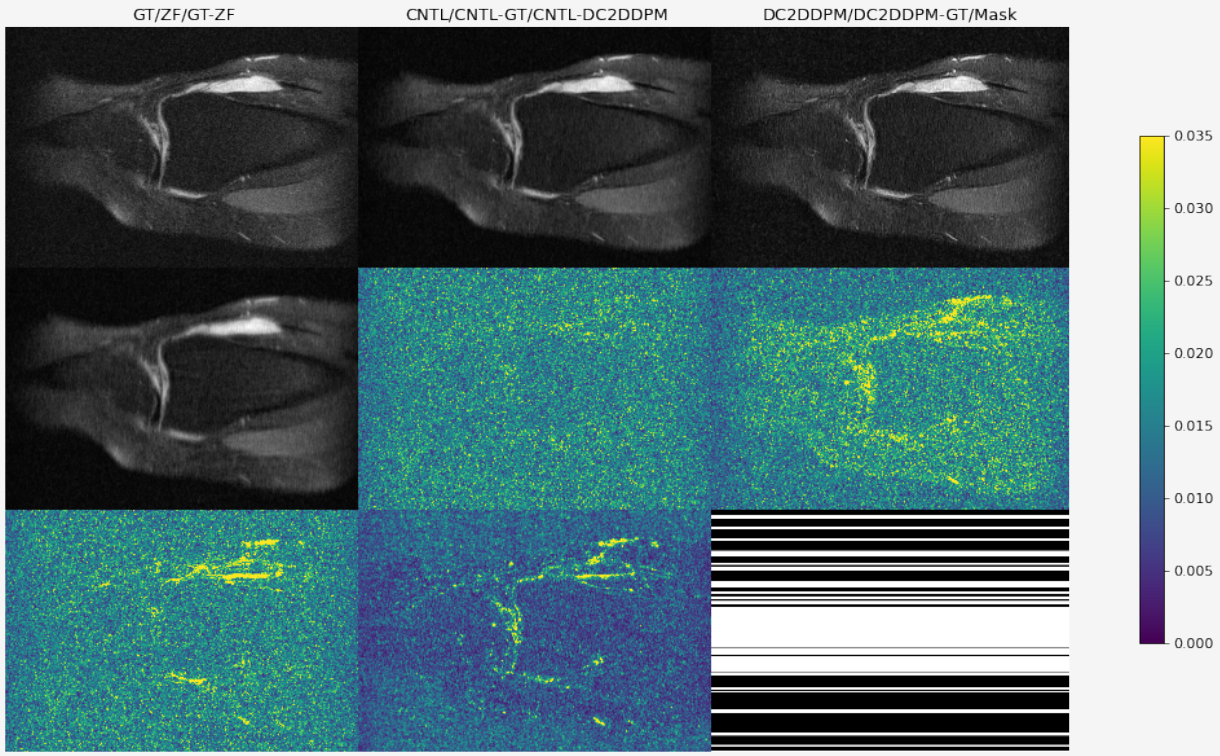

Figure 7.26: Reconstruction of x4 undersampled fastMRI single-coil knee image. (PSNR, SSIM) for CNTL and DC2DDPM: (33.8, 0.738), (33.2, 0.745).

certain image subsets from the UK BioBank study [150].

3. *Evaluation metrics for MRI reconstruction need further study* - The problem of image metrics for MRI reconstruction evaluation is closely related to the above point of the lack of gold-standard ground truth data. Issues with PSNR and SSIM as evaluation metrics have been highlighted in studies by [126, 206]. This has also been highlighted in the computer vision community [187, 46]. A recently proposed approach is the Video Multi-Method Assessment Fusion (VMAF) method [92]. This was developed as a way to evaluate compression methods for video data (TV and movies) by introducing opinion ratings of the end-user into the testing pipeline. The idea is to use a combination of different evaluation metrics such as VIF (Visual Information Fidelity; [25]) and DLM (Detail Loss Metric; [50]) in conjunction with an SVM to assign weights to each metric which is then used as a predictor for the image quality as viewed by an end-user.

In the case of the multi-coil data, we did not find significant improvement of the DDPM over

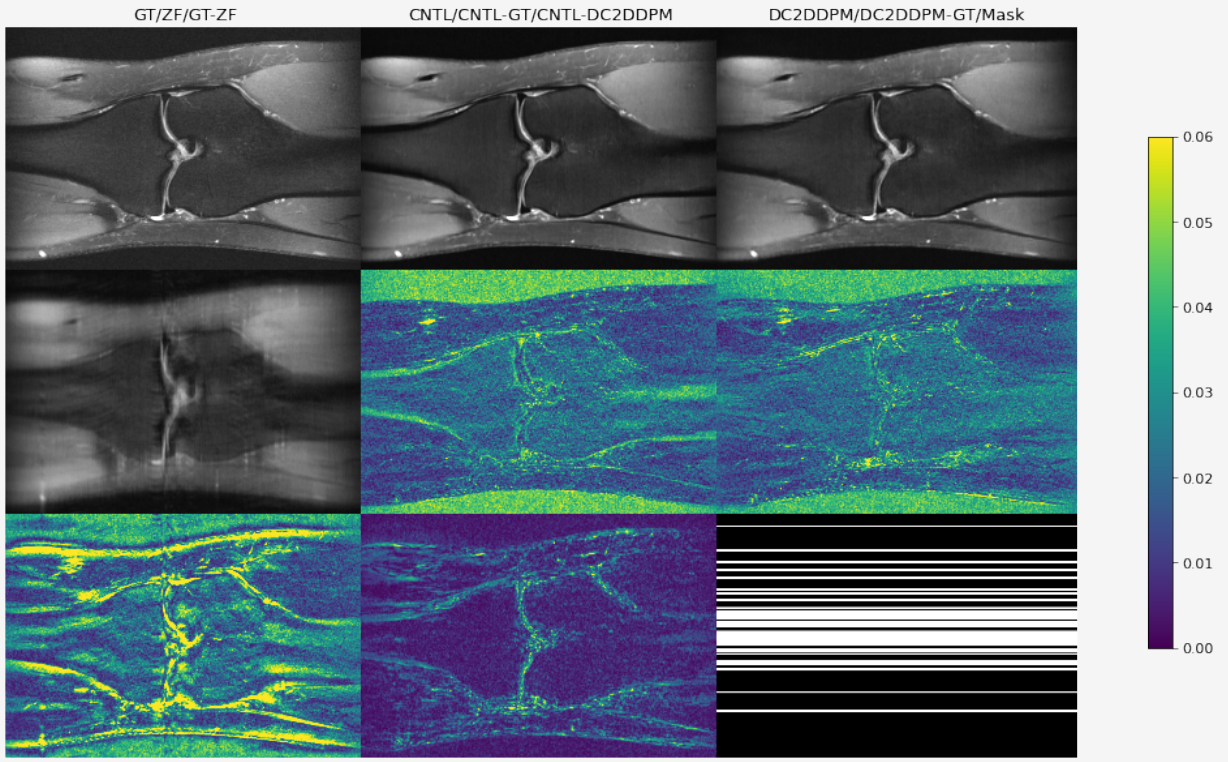

Figure 7.27: Reconstruction of x8 undersampled fastMRI multi-coil knee image. (PSNR, SSIM) for CNTL and DC2DDPM: (30.6, 0.795), (30.7, 0.800).

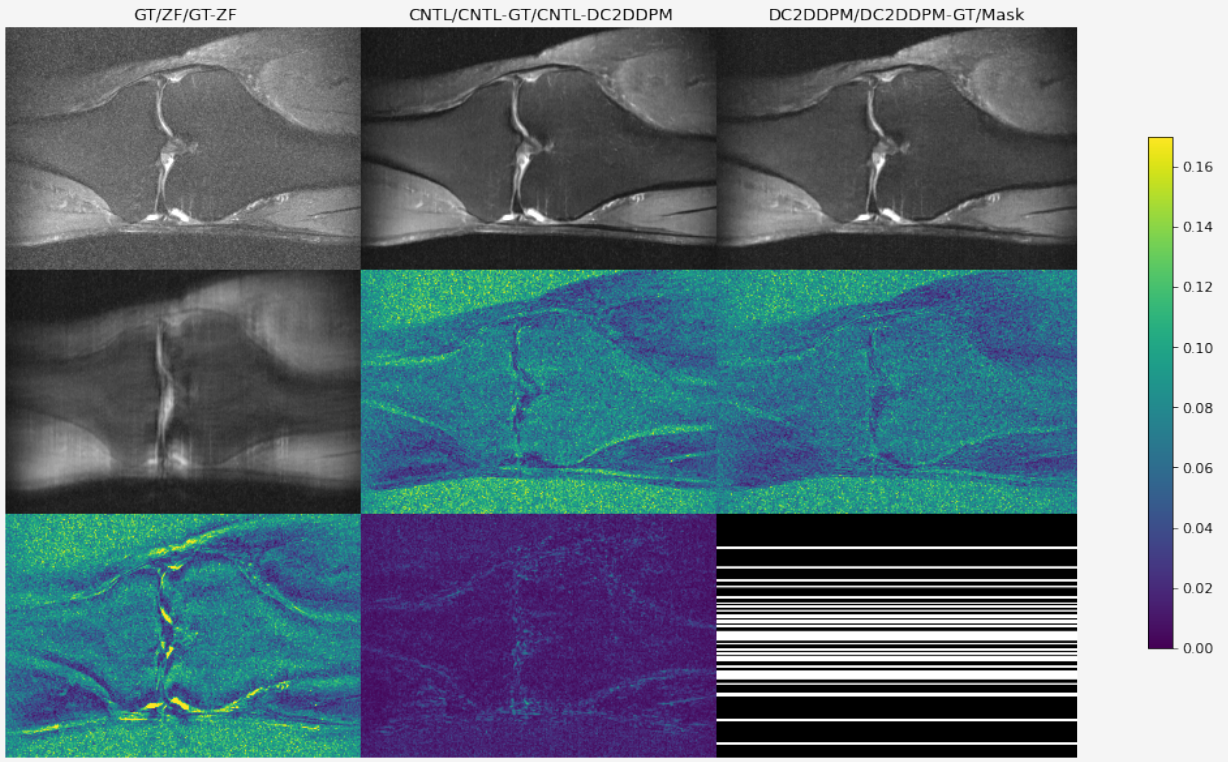

Figure 7.28: Reconstruction of x8 undersampled fastMRI multi-coil knee image. (PSNR, SSIM) for CNTL and DC2DDPM: (24.2, 0.643), (24.8, 0.650).

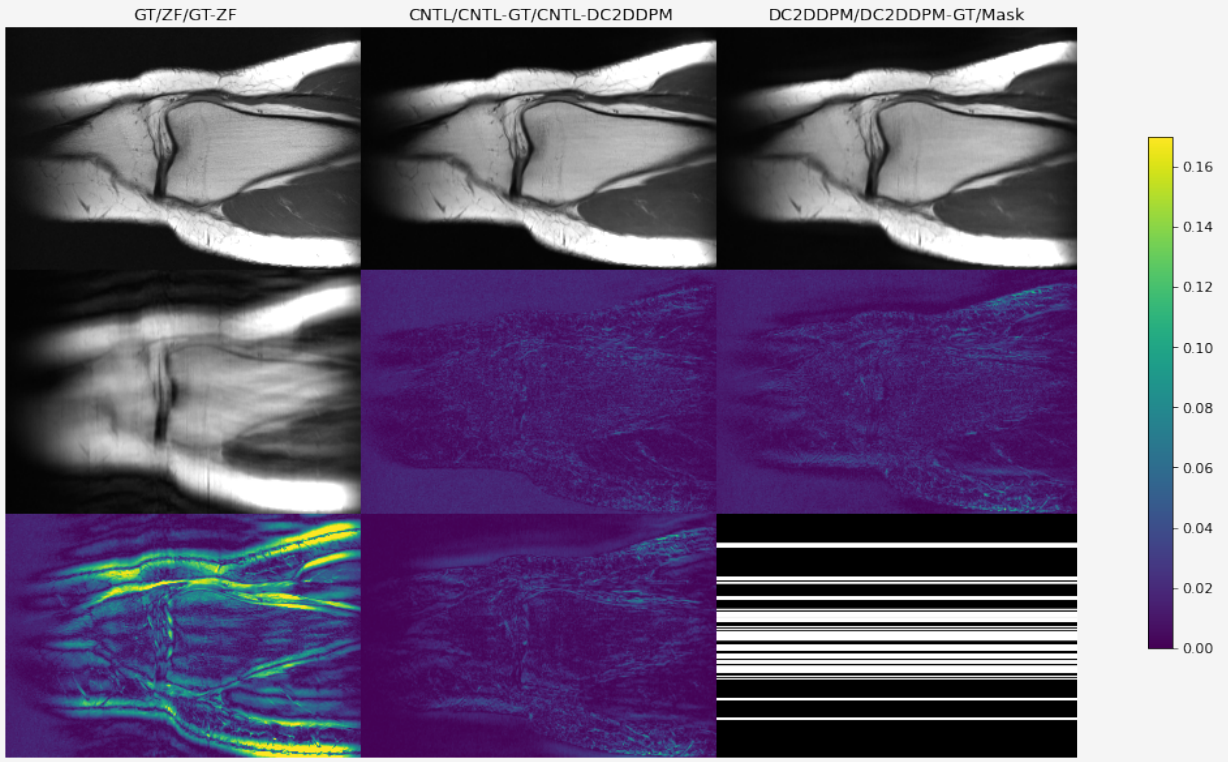

Figure 7.29: Reconstruction of x8 undersampled fastMRI multi-coil knee image. (PSNR, SSIM) for CNTL and DC2DDPM: (35.3, 0.862), (34.9, 0.880).

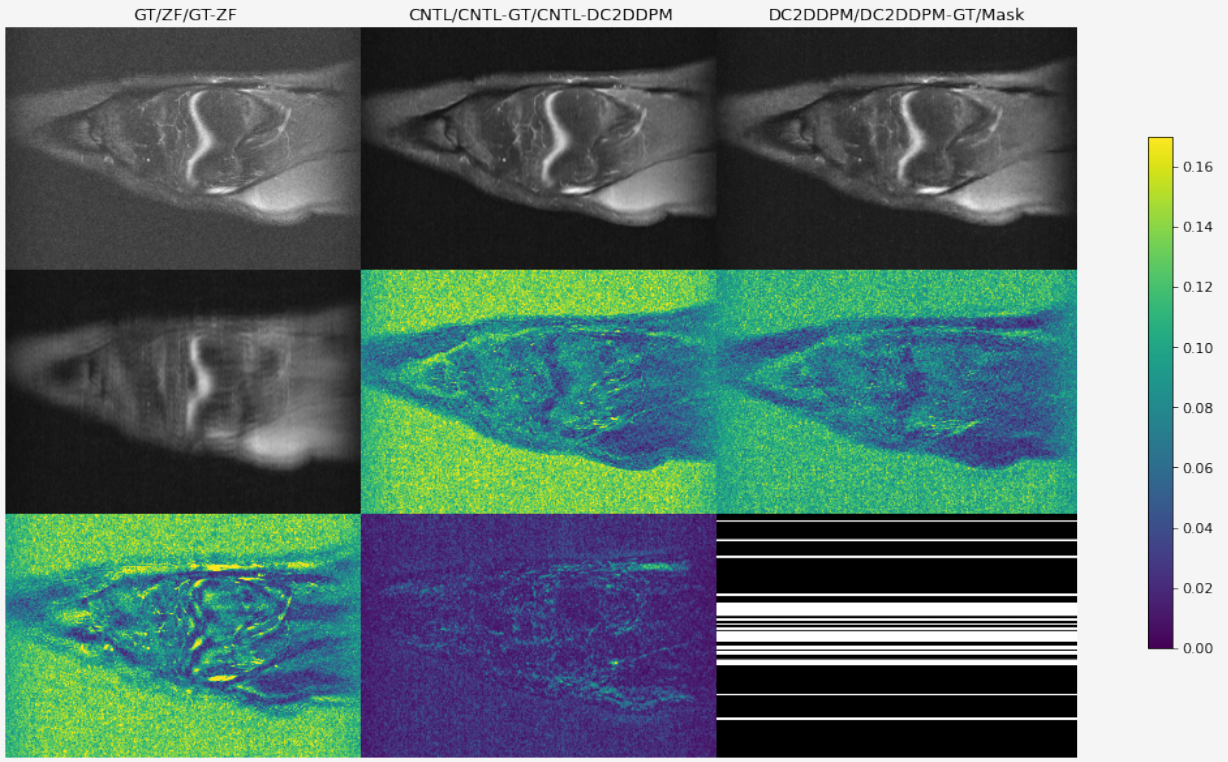

Figure 7.30: Reconstruction of x8 undersampled fastMRI multi-coil knee image. (PSNR, SSIM) for CNTL and DC2DDPM: (19.2, 0.472), (20.3, 0.488).

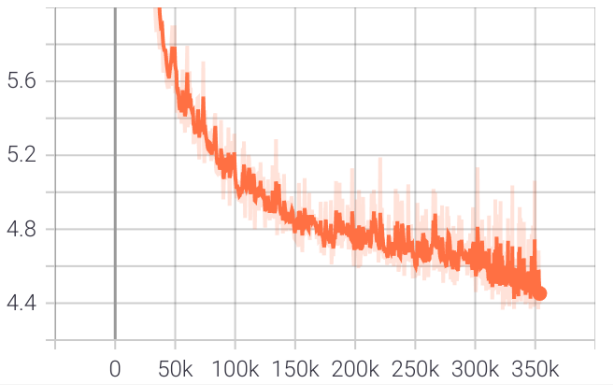

Figure 7.31: The plot of a validation metric against training step which ends at 350k after 20 days of training. The metric used is the L2 loss of the cascade's output prior to the DDPM. It is clear that the network has not fully converged.

the baseline. We hypothesise that this is mainly due to insufficient training. Training the single-coil DC2DDPM model took 20 times longer than that of the control. The multi-coil DC2DDPM - which handles more difficult data that the single coil case - did not converge even after 20 days of training on a 48GB NVIDIA RTX A6000. This is evident from Figure 7.31. Since training is very slow, a different approach is required to speed up training time to allow the network to fully converge.

Whilst the fidelity of the reconstructed multi-coil images are still competitive, further work is required to solve the problem of parallel MR image reconstruction with DC2DDPMs.

7.6.4 Conclusion

In this study on the fastMRI dataset, our proposed DDPM model was able to successfully learn a reconstruction model that seems to provide some interesting benefits over vanilla models. In particular, we found that our proposed model provided a sharper reconstruction and has potential to learn to mitigate imaging artefacts in the training data. Whilst the quantitative metrics for our proposal were not favourable, we attribute this to the lack of gold-standard ground truth data. Further investigation is required to ascertain the true quality of the image

reconstruction which may require opinion ratings from a selection of radiologists and clinicians[4]. For future work, we propose combining this work with that of the ME-DDPM from chapter 6 to perform reconstruction for cases where fully-sampled data is not available for cine acquisitions such as fetal CMR imaging.

7.7 Summary

In the first part of this study, we explored whether DDPMs could be used in the case where fully-sampled data was not available at training time (section 7.3). This was later extended to multiple reconstruction steps - not diffusion steps - which aim to directly solve MRI reconstruction optimisation problem (equation (7.1)) in section 7.4. Having successfully demonstrated the higher reconstruction fidelity of DDPMs we considered their performance in the presence of corrupted data. Using a non-Gaussian noise distribution to corrupt the image data in a scenario where fully-sampled data is available at training time, we found large performance gains by DDPMs over discriminative training (section 7.5). The efforts of all these studies were combined to learn a reconstruction model on the noisy fastMRI dataset without fully-sampled training data (retrospectively and uniquely applied undersampling masks). With this large real-world data, we found that our proposal seemed to generate higher fidelity reconstructions but performed worse on quantitative metrics. We leave it to future work to investigate whether these reconstructions are perceptually better and of a greater clinical relevance.

The main disadvantage of DDPMs is that due to their iterative nature, their inference time is extremely large (minimum of 30 seconds with our models if parallelised on a cluster of NVIDIA RTX A6000 GPUs) but there are works in the computer vision community which aim to reduce this burden [181].

[4]Note: we have created a visual comparison tool for this project to aid in the collection of opinion ratings which can be found here: `http://gavinseegoolam.co.uk/dc2ddpm/`. The 'Identifier' field can be anything (e.g. your name) and the login 'token' is 'phdthesis'.

Chapter 8

Conclusion

8.1 Summary of Thesis Achievements

In recent years, deep learning has shaken the world of medical imaging. During the course in which the work in this thesis was undertaken, there has been an exponentially increasing amount of published and accessible work in deep learning, computer vision and medicine, all of which will have even further impact on medical imaging. In this thesis, we contribute to this cause with the intention of creating smarter, more efficient tools for medical professionals.

Our biggest contribution comes in the form of accelerated MRI acquisition. Compressed sensing was firstly formally introduced to the problem of MRI reconstruction in 2006 [36]. During these 16 years, incremental studies have far advanced the state of MRI reconstruction. We have now introduced our own new methods, extending this initial breakthrough to realms which were not conceivable back in 2006. Our first development was the *motion exploiting convolutional neural network* or **ME-CNN**. This presents a way of performing dynamic MRI acquisitions 50 times faster than in the conventional approach. For higher fidelity image reconstructions, we present an adaption of this new method - ME-CNNv2 - which also had an interesting way to leverage abundant segmentation data in the UK BioBank dataset at training time. This was later dubbed the *motion-segmentation exploiting convolutional neural network*, the **MSE-CNN**.

Probabilistic deep learning is a subject that the research community is most familiar with for its ability to generate wonderful new images that do not exist via GANs and more recently, diffusion models [60, 128, 161, 184, 205]. We introduced probabilistic deep learning to the

problem of motion-based dynamic MR reconstruction in the form of the *motion exploiting denoising diffusion probabilistic model* or **ME-DDPM** with extremely promising results. We additionally proposed a new way to perform probabilistic deep learning unrolled reconstruction without the loss of fidelity or hallucination that usually occurs with GANs due to complications involved in adversarial training [162, 108, 84]. These developments were combined into an unsupervised framework called the *data consistent decomposed cascade denoising diffusion probabilistic model*, **DC2DDPM**. In this method, training could take place without the need for fully-sampled training data whilst simultaneously mitigating for data corruption in the data acquisition process. Accurate fully-sampled training data is often problematic for certain imaging scenarios such as fetal cardiac MR [156, 104] and corruption can take place in the form of noise, motion from the patient or other scanner imperfections. This was evaluated on the well-studied fastMRI knee with extremely promising results.

Stream-lining the diagnostic process is one of the ultimate aims of medical technology. This typically involves the production of an image which a radiologist uses to aid in the diagnosis of the underlying pathology. However, we questioned whether it's possible to acquire the MR data, skip the reconstruction phase and go straight to a diagnosis. Besides, in theory, the reconstruction process doesn't add any new information that wasn't already present in the acquired data. We demonstrated in our work in [124] that automated diagnostic report generation can take place at accelerations as high as $\times 8$ with noticeable drops in performance only occurring beyond this aggressive acceleration.

We outline the achievements of this thesis:

- Introduction of motion into the deep learning MRI reconstruction process - We presented the first deep learning method for generating reconstructions that exploit inter frame motion from accelerated acquisitions, the ME-CNN. We note that [190] introduce a kt-FOCUSS-like method for deep learning reconstruction however they require fully-sampled reference frames that hinder the acquisition process and its acceleration.

- Use of segmentation data for dynamic image reconstruction - We are the first to introduce

the concept of using segmentation data to ultimately generate better cine reconstructions. This combined the ideas of the ME-CNN and the work by [114].

- Provided insight into how to manipulate diffusion models for dynamic MRI to generate higher quality reconstructed cines
- Showed that much better performance can be achieved by diffusion models for MRI reconstruction by directly incorporating the unrolled MRI optimisation in the form of prior proximal-based cascades.
- Provided a method for training diffusion models without fully sampled acquisition data. We do not make use of fully-sampled targets and in this sense can be termed as probabilistic unsupervised MRI reconstruction
- Showed that diffusion models can be leveraged to mitigate data corruption without any modifications at training time even without ever seeing fully sampled data at training time. Since perfect, non-noisy data is never seen at training time, the setup is also unsupervised in another sense.
- Introduced a new method that decomposes diffusion models called DC2DDPM. This can be trained unsupervised on accelerated, noisy data in a single coil and multi-coil setting to generate corruption-mitigated reconstructions.
- Showed that automated diagnostic report generation can be achieved with undersampled MRI data

8.2 Future Work

Whilst this thesis provides the foundations for potential breakthroughs in the field of MRI acquisition, reconstruction and analysis, there is a plethora of further work that is to be conducted if the proposals are to be deployed in a clinical setting. Domain adaption is a key issue that is investigated in conjunction with the ME-CNN presented in this thesis. The study in [134] shows good generalisation of MRI reconstruction from one anatomical region to another.

However, for full clinical deployment, we require a better understanding of how artefacts in the imaging process are handled by such algorithms. For example, do pathology hallucinations occur in the presence of magnetic susceptibility artefacts? Further work is required to ensure such issues in the MRI reconstruction process do not occur or are identifiable using statistical techniques. In particular, this should be studied for our work on diffusion models, where there is currently no existing literature pertaining to this. A notable mention is Stein's unbiased risk estimator (SURE) which has been a focus of recent work in the field of MRI reconstruction [165]. In brief, SURE provides access to the accuracy of the reconstruction using theoretical guarantees of the devised model [157].

Whilst the proposed ME-CNN exploits motion for the image reconstruction process, there isn't a mechanism to correct intra-frame motion. This can be considered from two perspectives: 1) Mitigating for the problem of intra-frame motion in the target of the ME-CNN at training time 2) Intra-frame motion in the undersampled input data from rapid patient movement during scanning. This is particularly of interest in fetal cardiac MR where the issue of motion is severe and parallel imaging is typically employed. One possible method to partially alleviate this problem with the ME-CNN is to train the network in an unsupervised fashion with real highly undersampled acquisitions rather than motion-corrupted fully-sampled data. The main complication that arises is training the motion estimators in an unsupervised fashion, however one hypothesis is that this is possible by consecutively training the ME-CNN reconstruction network and then training the motion network to use the end-reconstruction of the ME-CNN. During training, the optimisation landscape will constantly evolve - as the motion estimate gets better and better, so does the reconstruction.

In this thesis, we focused on Cartesian acquisitions however, for example in brain MRI, it is well noted that radial trajectories lead to better reconstructions due to motion robustness and greater incoherence for compressed sensing MRI [129, 18, 15]. The problem of deep learning reconstruction for non-Cartesian acquisitions has previously been studied [139, 157, 174] and we propose that this is an appropriate extension to the work presented in this thesis. We also focus mainly on the proximal-based approach in this thesis where the proximal mapping is off-loaded to a neural network. However, gradient descent variants for all work in this thesis

are possible and should be investigated further.

Probabilistic deep learning for MRI reconstruction has vastly inferior literature compared with its deterministic counter-part. Whilst this thesis produces methods that help bridge this gap, the ME-DDPM presented in this thesis has a clear disadvantage in that the motion estimate is not trained end-to-end with the reconstruction model. Furthermore, our work for the DC2DDPM shows a way in which the ME-DDPM can be further improved by incorporation of cascading. For future work, we propose incorporating a motion estimator in the cascades, similar to the ME-CNN. At test time, the gradually refined motion estimate can be used in the same way as in the ME-DDPM, to drive the latent representation closer towards the true posterior.

8.3 Final Remark

In summary, whilst this thesis has advanced the community's state of knowledge on the topic of MRI reconstruction, there is a lot of future research required to advance towards clinical deployment. Motion exploitation offers a lot of potential and there is now an appropriate approach to probabilistic modelling for MRI reconstruction. We hope to continue to grow this research going forward, working closely with other institutions, radiologists and clinicians.

Chapter A

Supplementary Material

A.1 Supplementary Material 1

Please find Supplementary Material 1 at `http://gavinseegoolam.co.uk/wp/thesis_1-1/`.

This is a video clip which depicts several cardiac cines in an accelerated setting. This is an example of a retrospectively x16 accelerated acquisition which is then reconstructed with zero-filling (left). The middle cine is the data sharing cine with a depth of 5 [93]. The right cine is the x-DC-MAC generated using a crude motion estimate obtained using from an optical-flow based autoencoder-like network (not a U-net). We use a crude motion estimate (and thus an autoencoder-like network, not a U-net) to demonstrate that a perfect motion estimate is not required to see the huge advantages of x-DC-MAC compared to DS found in DC-CNN.

A.2 Supplementary Materials 2a-6c

These can be found at the following URLs:

- Supplementary Material 2a: http://gavinseegoolam.co.uk/wp/thesis_1-6a/.
- Supplementary Material 2b: http://gavinseegoolam.co.uk/wp/thesis_1-6b/.
- Supplementary Material 2c: http://gavinseegoolam.co.uk/wp/thesis_1-6c/.
- Supplementary Material 2d: http://gavinseegoolam.co.uk/wp/thesis_1-6d/.
- Supplementary Material 3a: http://gavinseegoolam.co.uk/wp/thesis_1-7a/.
- Supplementary Material 3b: http://gavinseegoolam.co.uk/wp/thesis_1-7b/.
- Supplementary Material 3c: http://gavinseegoolam.co.uk/wp/thesis_1-7c/.
- Supplementary Material 3d: http://gavinseegoolam.co.uk/wp/thesis_1-7d/.
- Supplementary Material 4a: http://gavinseegoolam.co.uk/wp/thesis_1-8a/.
- Supplementary Material 4b: http://gavinseegoolam.co.uk/wp/thesis_1-8b/.
- Supplementary Material 4c: http://gavinseegoolam.co.uk/wp/thesis_1-8c/.
- Supplementary Material 4d: http://gavinseegoolam.co.uk/wp/thesis_1-8d/.
- Supplementary Material 5a: http://gavinseegoolam.co.uk/wp/thesis_1-9a/.
- Supplementary Material 5b: http://gavinseegoolam.co.uk/wp/thesis_1-9b/.
- Supplementary Material 5c: http://gavinseegoolam.co.uk/wp/thesis_1-9c/.
- Supplementary Material 5d: http://gavinseegoolam.co.uk/wp/thesis_1-9d/.
- Supplementary Material 6a: http://gavinseegoolam.co.uk/wp/thesis_1-10a/.
- Supplementary Material 6b: http://gavinseegoolam.co.uk/wp/thesis_1-10b/.
- Supplementary Material 6c: http://gavinseegoolam.co.uk/wp/thesis_1-10c/.

Bibliography

[1] Brian DO Anderson. "Reverse-time diffusion equation models". In: *Stochastic Processes and their Applications* 12.3 (1982), pp. 313–326.

[2] Kurt O Jörnsten, Mikael Näsberg, and Per A Smeds. *Variable splitting: A new Lagrangean relaxation approach to some mathematical programming models*. Universitetet i Linköping/Tekniska Högskolan i Linköping. Department of ..., 1985.

[3] Monique Guignard and Siwhan Kim. "Lagrangean decomposition: A model yielding stronger Lagrangean bounds". In: *Mathematical programming* 39.2 (1987), pp. 215–228.

[4] Alfred L Horowitz and Alfred L Horowitz. *MRI physics for radiologists*. Springer, 1992.

[5] Jerry L Prince and Elliot R McVeigh. "Motion estimation from tagged MR image sequences". In: *IEEE transactions on medical imaging* 11.2 (1992), pp. 238–249.

[6] Leonid I Rudin, Stanley Osher, and Emad Fatemi. "Nonlinear total variation based noise removal algorithms". In: *Physica D: nonlinear phenomena* 60.1-4 (1992), pp. 259–268.

[7] Allen D Elster. "Gradient-echo MR imaging: techniques and acronyms." In: *Radiology* 186.1 (1993), pp. 1–8.

[8] Albert Macovski and Steven Conolly. "Novel approaches to low-cost MRI". In: *Magnetic resonance in medicine* 30.2 (1993), pp. 221–230.

[9] David A Feinberg, Neil M Rofsky, and Glyn Johnson. "Multiple breath-hold averaging (mba) method for increased snr in abdominal mri". In: *Magnetic resonance in medicine* 34.6 (1995), pp. 905–909.

[10] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[11] Winfried Lohmiller and Jean-Jacques E Slotine. "On contraction analysis for non-linear systems". In: *Automatica* 34.6 (1998), pp. 683–696.

[12] Thomas William Redpath. "Signal-to-noise ratio in MRI." In: *The British journal of radiology* 71.847 (1998), pp. 704–707.

[13] Jan Sijbers et al. "Maximum-likelihood estimation of Rician distribution parameters". In: *IEEE Transactions on Medical Imaging* 17.3 (1998), pp. 357–361.

[14] Gene H Golub, Per Christian Hansen, and Dianne P O'Leary. "Tikhonov regularization and total least squares". In: *SIAM journal on matrix analysis and applications* 21.1 (1999), pp. 185–194.

[15] James G Pipe. "Motion correction with PROPELLER MRI: application to head motion and free-breathing cardiac imaging". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 42.5 (1999), pp. 963–969.

[16] Klaas P Pruessmann et al. "SENSE: sensitivity encoding for fast MRI". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 42.5 (1999), pp. 952–962.

[17] Daniel Rueckert et al. "Nonrigid registration using free-form deformations: application to breast MR images". In: *IEEE transactions on medical imaging* 18.8 (1999), pp. 712–721.

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

[19] Andrew Ng and Michael Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". In: *Advances in neural information processing systems* 14 (2001).

[20] Martin J Wainwright, Eero P Simoncelli, and Alan S Willsky. "Random cascades on wavelet trees and their use in analyzing and modeling natural images". In: *Applied and Computational Harmonic Analysis* 11.1 (2001), pp. 89–123.

[21] Mark A Griswold et al. "Generalized autocalibrating partially parallel acquisitions (GRAPPA)". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 47.6 (2002), pp. 1202–1210.

[22] Kishore Papineni et al. "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2002, pp. 311–318.

[23] Chin-Yew Lin and Eduard Hovy. "Automatic evaluation of summaries using n-gram co-occurrence statistics". In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 2003, pp. 150–157.

[24] Jeffrey Tsao, Peter Boesiger, and Klaas P Pruessmann. "k-t BLAST and k-t SENSE: dynamic MRI with high frame rate exploiting spatiotemporal correlations". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 50.5 (2003), pp. 1031–1042.

[25] H. R. Sheikh and A. C. Bovik. "Image information and visual quality". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3. May 2004, pp. iii–709. DOI: 10.1109/ICASSP.2004.1326643.

[26] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[27] PG Batchelor et al. "Matrix description of general motion correction applied to multishot images". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 54.5 (2005), pp. 1273–1280.

[28] Michael Lustig et al. "Application of compressed sensing for rapid MR imaging". In: *SPARS,(Rennes, France)* (2005).

[29] Stefan Roth and Michael J Black. "Fields of experts: A framework for learning image priors". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 2. IEEE. 2005, pp. 860–867.

[30] Emmanuel J Candès, Justin Romberg, and Terence Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on information theory* 52.2 (2006), pp. 489–509.

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

[32] Francois Rousseau et al. "Registration-based approach for reconstruction of high-resolution in utero fetal MR brain images". In: *Academic radiology* 13.9 (2006), pp. 1072–1081.

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

[34] Kostadin Dabov et al. "Image denoising by sparse 3-D transform-domain collaborative filtering". In: *IEEE Transactions on image processing* 16.8 (2007), pp. 2080–2095.

[35] Hong Jung, Jong Chul Ye, and Eung Yeop Kim. "Improved k-t BLAST and k-t SENSE using FOCUSS". In: *Physics in Medicine and Biology* 52.11 (2007), pp. 3201–3226. ISSN: 00319155. DOI: 10.1088/0031-9155/52/11/018.

[36] Michael Lustig, David Donoho, and John M Pauly. "Sparse MRI: The application of compressed sensing for rapid MR imaging". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 58.6 (2007), pp. 1182–1195.

[37] Dominique Béréziat and Isabelle L Herlin. "Solving ill-posed Image Processing problems using Data Assimilation. Application to optical flow". PhD thesis. INRIA, 2008.

[38] Pierre Courrieu. "Fast computation of Moore-Penrose inverse matrices". In: *arXiv preprint arXiv:0804.4809* (2008).

[39] Germana Landi, Elena Loli Piccolomini, and Fabiana Zama. "A total variation-based reconstruction method for dynamic MRI". In: *Computational and Mathematical Methods in Medicine* 9.1 (2008), pp. 69–80.

[40] Michael Lustig et al. "Compressed sensing MRI". In: *IEEE signal processing magazine* 25.2 (2008), pp. 72–82.

[41] Amir Beck and Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM journal on imaging sciences* 2.1 (2009), pp. 183–202.

[42] Hong Jung et al. "K-t FOCUSS: A general compressed sensing framework for high resolution dynamic MRI". In: *Magnetic Resonance in Medicine* 61.1 (2009), pp. 103–116. ISSN: 15222594. DOI: 10.1002/mrm.21757.

[43] Quang-Cuong Pham, Nicolas Tabareau, and Jean-Jacques Slotine. "A contraction theory approach to stochastic incremental stability". In: *IEEE Transactions on Automatic Control* 54.4 (2009), pp. 816–820.

[44] Manya V Afonso, José M Bioucas-Dias, and Mário AT Figueiredo. "Fast image recovery using variable splitting and constrained optimization". In: *IEEE transactions on image processing* 19.9 (2010), pp. 2345–2356.

[45] Catherine Foot, Chris Naylor, Candace Imison, et al. "The quality of GP diagnosis and referral". In: (2010).

[46] Alain Hore and Djemel Ziou. "Image quality metrics: PSNR vs. SSIM". In: *2010 20th international conference on pattern recognition*. IEEE. 2010, pp. 2366–2369.

[47] Florian Knoll et al. "Total Generalized Variation (TGV) for MRI". In: *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 18. 2010, p. 4855.

[48] Xiaobo Qu et al. "Iterative thresholding compressed sensing MRI based on contourlet transform". In: *Inverse Problems in Science and Engineering* 18.6 (2010), pp. 737–758.

[49] Saiprasad Ravishankar and Yoram Bresler. "MR image reconstruction from highly undersampled k-space data by dictionary learning". In: *IEEE transactions on medical imaging* 30.5 (2010), pp. 1028–1041.

[50] Songnan Li et al. “Image quality assessment by separately evaluating detail losses and additive impairments”. In: *IEEE Transactions on Multimedia* 13.5 (2011), pp. 935–949.

[51] Sajan Goud Lingala et al. “Accelerated dynamic MRI exploiting sparsity and low-rank structure: kt SLR”. In: *IEEE transactions on medical imaging* 30.5 (2011), pp. 1042–1054.

[52] Caroline Petitjean and Jean-Nicolas Dacher. “A review of segmentation methods in short axis cardiac MR images”. In: *Medical image analysis* 15.2 (2011), pp. 169–184.

[53] João Carlos Alves Barata and Mahir Saleh Hussein. “The Moore–Penrose pseudoinverse: A tutorial review of the theory”. In: *Brazilian Journal of Physics* 42.1 (2012), pp. 146–165.

[54] Jose Caballero, Daniel Rueckert, and Joseph V Hajnal. “Dictionary learning and time sparsity in dynamic MRI”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2012, pp. 256–263.

[55] NJ Levell, AM Penart-Lanau, and JJ Garioch. “Introduction of intermediate care dermatology services in Norfolk, England was followed by a 67% increase in referrals to the local secondary care dermatology department”. In: *British Journal of Dermatology* 167.2 (2012), pp. 443–445.

[56] Sairam Geethanath et al. “Compressed sensing MRI: a review”. In: *Critical Reviews™ in Biomedical Engineering* 41.3 (2013).

[57] UK Biobank. *About uk biobank*. 2014.

[58] Robert W Brown et al. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

[59] Jose Caballero et al. “Application-driven MRI: joint reconstruction and segmentation from undersampled MRI data”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2014, pp. 106–113.

[60] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. DOI: 10.48550/ARXIV.1406.2661. URL: https://arxiv.org/abs/1406.2661.

[61] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[62] Martin Uecker et al. "ESPIRiT—an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA". In: *Magnetic resonance in medicine* 71.3 (2014), pp. 990–1001.

[63] Ivana Despotović, Bart Goossens, and Wilfried Philips. "MRI segmentation of the human brain: challenges, methods, and applications". In: *Computational and mathematical methods in medicine* 2015 (2015).

[64] Oren N Jaspan, Roman Fleysher, and Michael L Lipton. "Compressed sensing MRI: a review of the clinical literature". In: *The British journal of radiology* 88.1056 (2015), p. 20150487.

[65] Ritse M Mann et al. "Breast MRI: EUSOBI recommendations for women's information". In: *European radiology* 25.12 (2015), pp. 3669–3678.

[66] Junhua Mao et al. "Deep captioning with multimodal recurrent neural networks (m-rnn)". In: *ICLR* (2015).

[67] Steffen E Petersen et al. "UK Biobank's cardiovascular magnetic resonance protocol". In: *Journal of cardiovascular magnetic resonance* 18.1 (2015), pp. 1–7.

[68] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[69] Cathie Sudlow et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS medicine* 12.3 (2015), e1001779.

[70] Oriol Vinyals et al. "Show and tell: A neural image caption generator". In: *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE. 2015, pp. 3156–3164.

[71] Kelvin Xu et al. "Show, attend and tell: Neural image caption generation with visual attention". In: *International Conference on Machine Learning*. 2015, pp. 2048–2057.

[72] Jiawen Yao et al. "Accelerated dynamic MRI reconstruction with total variation and nuclear norm regularization". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 635–642.

[73] Zhifang Zhan et al. "Fast multiclass dictionaries learning with geometrical directions in MRI reconstruction". In: *IEEE Transactions on biomedical engineering* 63.9 (2015), pp. 1850–1861.

[74] Aria Ahmadi and Ioannis Patras. "Unsupervised convolutional neural networks for motion estimation". In: *2016 IEEE international conference on image processing (ICIP)*. IEEE. 2016, pp. 1629–1633.

[75] Vinit Baliyan et al. "Diffusion weighted imaging: technique and applications". In: *World journal of radiology* 8.9 (2016), p. 785.

[76] Jose Caballero et al. "Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation". In: *arXiv e-prints*, arXiv:1611.05250 (Oct. 2016), arXiv:1611.05250. arXiv: `1611.05250 [cs.CV]`.

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

[78] Lucilio Cordero Grande et al. "3D motion corrected SENSE reconstruction for multishot multislice MRI". In: *conference abstracts of ISMRM*. 2016.

[79] C. David Preston. *MRI Basics*. 2016. URL: `https://case.edu/med/neurology/NR/MRI%5C%20Basics.htm`.

[80] Hoo-Chang Shin et al. "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2497–2506.

[81] Baochen Sun, Jiashi Feng, and Kate Saenko. "Return of frustratingly easy domain adaptation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1. 2016.

[82] Jian Sun, Huibin Li, Zongben Xu, et al. "Deep ADMM-Net for compressive sensing MRI". In: *Advances in neural information processing systems* 29 (2016).

[83] Hang Zhao et al. "Loss functions for image restoration with neural networks". In: *IEEE Transactions on computational imaging* 3.1 (2016), pp. 47–57.

[84] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

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

[86] Ishaan Gulrajani et al. "Improved training of wasserstein gans". In: *Advances in neural information processing systems* 30 (2017).

[87] Eddy Ilg et al. "Flownet 2.0: Evolution of optical flow estimation with deep networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2462–2470.

[88] Chun-Liang Li et al. "Mmd gan: Towards deeper understanding of moment matching network". In: *Advances in neural information processing systems* 30 (2017).

[89] Brett Lorraine and Charles Lott. *MRI Heart (Cardiac MRI)*. 2017. URL: `https://www.insideradiology.com.au/cardiac-mri/` (visited on 07/02/2022).

[90] Tim Meinhardt et al. "Learning proximal operators: Using denoising networks for regularizing inverse imaging problems". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1781–1790.

[91] Andreas Merrem et al. "Rapid diffusion-weighted magnetic resonance imaging of the brain without susceptibility artifacts: Single-shot STEAM with radial undersampling and iterative reconstruction". In: *Investigative radiology* 52.7 (2017), pp. 428–433.

[92] Netflix. *Toward A Practical Perceptual Video Quality Metric*. 2017. URL: `https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652`.

[93] Jo Schlemper et al. "A deep cascade of convolutional neural networks for dynamic MR image reconstruction". In: *IEEE transactions on Medical Imaging* 37.2 (2017), pp. 491–503.

[94] Hessam Sokooti et al. "Nonrigid image registration using multi-scale 3D convolutional neural networks". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2017, pp. 232–239.

[95] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[96] Jakob Weiss et al. "Feasibility of accelerated simultaneous multislice diffusion-weighted MRI of the prostate". In: *Journal of Magnetic Resonance Imaging* 46.5 (2017), pp. 1507–1515.

[97] Wenchuan Wu and Karla L Miller. "Image formation in diffusion MRI: a review of recent technical developments". In: *Journal of Magnetic Resonance Imaging* 46.3 (2017), pp. 646–662.

[98] Guang Yang et al. "DAGAN: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction". In: *IEEE transactions on medical imaging* 37.6 (2017), pp. 1310–1321.

[99] Manzil Zaheer et al. "GAN connoisseur: Can GANs learn simple 1D parametric distributions". In: *Proceedings of the 31st Conference on Neural Information Processing Systems*. 2017, pp. 1–6.

[100] Kai Zhang et al. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising". In: *IEEE transactions on image processing* 26.7 (2017), pp. 3142–3155.

[101] Zizhao Zhang et al. "Mdnet: A semantically and visually interpretable medical image diagnosis network". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6428–6436.

[102] Jonas Adler and Ozan Öktem. "Deep bayesian inversion". In: *arXiv preprint arXiv:1811.05910* (2018).

[103] Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. "MoDL: Model-based deep learning architecture for inverse problems". In: *IEEE transactions on medical imaging* 38.2 (2018), pp. 394–405.

[104] Joshua FP van Amerom et al. "Fetal cardiac cine imaging using highly accelerated dynamic MRI with retrospective motion correction and outlier rejection". In: *Magnetic resonance in medicine* 79.1 (2018), pp. 327–338.

[105] Orli G Bahcall. "UK Biobank—a new era in genomic medicine". In: *Nature reviews genetics* 19.12 (2018), pp. 737–737.

[106] Arantxa Casanova et al. *On the iterative refinement of densely connected representation levels for semantic segmentation*. 2018. arXiv: `1804.11332 [cs.CV]`.

[107] Alexander Ciritsis et al. "Accelerated diffusion-weighted imaging for lymph node assessment in the pelvis applying simultaneous multislice acquisition: a healthy volunteer study". In: *Medicine* 97.32 (2018).

[108] Antonia Creswell et al. "Generative adversarial networks: An overview". In: *IEEE signal processing magazine* 35.1 (2018), pp. 53–65.

[109] Ahmet Mesrur Halefoglu and David Mark Yousem. "Susceptibility weighted imaging: clinical applications and future directions". In: *World journal of radiology* 10.4 (2018), p. 30.

[110] Kerstin Hammernik et al. "Learning a variational network for reconstruction of accelerated MRI data". In: *Magnetic resonance in medicine* 79.6 (2018), pp. 3055–3071.

[111] Baoyu Jing, Pengtao Xie, and Eric Xing. "On the Automatic Generation of Medical Imaging Reports". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 2577–2586.

[112] Anna Nowogrodzki. "The world's strongest MRI machines are pushing human imaging to new limits". In: *Nature* 563.7732 (2018), pp. 24–27.

[113] Chen Qin et al. "Convolutional recurrent neural networks for dynamic MR image reconstruction". In: *IEEE transactions on medical imaging* 38.1 (2018), pp. 280–290.

[114] Chen Qin et al. "Joint learning of motion estimation and segmentation for cardiac MR image sequences". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 472–480.

[115] Edward T Reehorst and Philip Schniter. "Regularization by denoising: Clarifications and new interpretations". In: *IEEE transactions on computational imaging* 5.1 (2018), pp. 52–67.

[116] Jo Schlemper et al. "Cardiac MR segmentation from undersampled k-space using deep latent representation learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 259–267.

[117] Maximilian Seitzer et al. "Adversarial and perceptual refinement for compressed sensing MRI reconstruction". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2018, pp. 232–240.

[118] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. "Deep image prior". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9446–9454.

[119] Yuan Xue et al. "Multimodal recurrent model with attention for automated radiology report generation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 457–466.

[120] Jure Zbontar et al. "fastMRI: An open dataset and benchmarks for accelerated MRI". In: *arXiv preprint arXiv:1811.08839* (2018).

[121] Bo Zhu et al. "Image reconstruction by domain-transform manifold learning". In: *Nature* 555.7697 (2018), pp. 487–492.

[122] Guha Balakrishnan et al. "VoxelMorph: a learning framework for deformable medical image registration". In: *IEEE transactions on medical imaging* 38.8 (2019), pp. 1788–1800.

[123] Jinming Duan et al. "VS-Net: Variable splitting network for accelerated parallel MRI reconstruction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 713–722.

[124] Aydan Gasimova. "Automated Enriched Medical Concept Generation for Chest X-ray Images". In: *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, 2019, pp. 83–92.

[125] Vahid Ghodrati et al. "MR image reconstruction using deep learning: evaluation of network structure and loss functions". In: *Quantitative imaging in medicine and surgery* 9.9 (2019), p. 1516.

[126] Kerstin Hammernik et al. "Sigma-net: Systematic Evaluation of Iterative Deep Neural Networks for Fast Parallel MR Image Reconstruction". In: *arXiv preprint arXiv:1912.09278* (2019).

[127] Qiaoying Huang et al. "FR-Net: Joint reconstruction and segmentation in compressed sensing cardiac MRI". In: *International conference on functional imaging and modeling of the heart*. Springer. 2019, pp. 352–360.

[128] Tero Karras, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.

[129] Carole Lazarus et al. "SPARKLING: variable-density k-space filling curves for accelerated T2*-weighted MRI". In: *Magnetic resonance in medicine* 81.6 (2019), pp. 3643–3661.

[130] Guanxiong Liu et al. "Clinically accurate chest x-ray report generation". In: *Machine Learning for Healthcare Conference*. PMLR. 2019, pp. 249–269.

[131] Nicholas McKibben. *k-t BLAST*. 2019. URL: `https://github.com/mckib2/ktblast`.

[132] Rosa-Marıa Menchón-Lara et al. "Reconstruction techniques for cardiac cine MRI". In: *Insights into imaging* 10.1 (2019), pp. 1–16.

[133] NVIDIA. *TensorFlow Determinism/Framework Determinism*. 2019. URL: `https://github.com/NVIDIA/framework-determinism`.

[134] Cheng Ouyang et al. "Generalising deep learning MRI reconstruction across different domains". In: *arXiv preprint arXiv:1902.10815* (2019).

[135] Duncan Riach. *TensorFlow Determinism*. 2019. URL: `https://bit.ly/dl-determinism-slides-v3`.

[136] Simo Särkkä and Arno Solin. *Applied stochastic differential equations*. Vol. 10. Cambridge University Press, 2019.

[137] Elisabeth Sartoretti et al. "Reduction of procedure times in routine clinical practice with Compressed SENSE magnetic resonance imaging technique". In: *PLoS One* 14.4 (2019), e0214887.

[138] Jo Schlemper et al. "Data consistency networks for (calibration-less) accelerated parallel MR image reconstruction". In: *arXiv preprint arXiv:1909.11795* (2019).

[139] Jo Schlemper et al. "Nonuniform variational network: deep learning for accelerated nonuniform MR image reconstruction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 57–64.

[140] Jo Schlemper et al. "Sigma-net: Ensembled Iterative Deep Neural Networks for Accelerated Parallel MR Image Reconstruction". In: *arXiv preprint arXiv:1912.05480* (2019).

[141] Gavin Seegoolam et al. "Exploiting Motion for Deep Learning Reconstruction of Extremely-Undersampled Dynamic MRI". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 704–712.

[142] T. Smyth. *Nyquist Sampling Theorem*. 2019.

[143] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in Neural Information Processing Systems* 32 (2019).

[144] Liyan Sun et al. "Joint CS-MRI reconstruction and segmentation with a unified deep network". In: *International conference on information processing in medical imaging*. Springer. 2019, pp. 492–504.

[145] Dirk Voit, Oleksandr Kalentev, and Jens Frahm. "Body coil reference for inverse reconstructions of multi-coil data—the case for real-time MRI". In: *Quantitative Imaging in Medicine and Surgery* 9.11 (2019), p. 1815.

[146] Jianbo Yuan et al. "Automatic Radiology Report Generation based on Multi-view Image Fusion and Medical Concept Enrichment". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 721–729.

[147] Zizhao Zhang et al. "Reducing Uncertainty in Undersampled MRI Reconstruction with Active Acquisition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2049–2058.

[148] Shengyu Zhao et al. "Recursive cascaded networks for unsupervised medical image registration". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 10600–10610.

[149] Rizwan Ahmad et al. "Plug-and-play methods for magnetic resonance imaging: Using denoisers for image recovery". In: *IEEE signal processing magazine* 37.1 (2020), pp. 105–116.

[150] Wenjia Bai et al. "A population-based phenome-wide association study of cardiac and aortic structure and function". In: *Nature medicine* 26.10 (2020), pp. 1654–1662.

[151] Wanyu Bian, Yunmei Chen, and Xiaojing Ye. "Deep parallel MRI reconstruction network without coil sensitivities". In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2020, pp. 17–26.

[152] Chen Chen et al. "Deep learning for cardiac image segmentation: a review". In: *Frontiers in Cardiovascular Medicine* 7 (2020), p. 25.

[153] Dongdong Chen, Mike E Davies, and Mohammad Golbabaee. "Compressive mr fingerprinting reconstruction with neural proximal gradient iterations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 13–22.

[154] Elizabeth K Cole et al. "Unsupervised MRI reconstruction with generative adversarial networks". In: *arXiv preprint arXiv:2008.13065* (2020).

[155] Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.

[156] Su-Zhen Dong et al. "Fetal cardiac MRI: a single center experience over 14-years on the potential utility as an adjunct to fetal technically inadequate echocardiography". In: *Scientific Reports* 10.1 (2020), pp. 1–10.

[157] Vineet Edupuganti et al. "Uncertainty quantification in deep MRI reconstruction". In: *IEEE Transactions on Medical Imaging* 40.1 (2020), pp. 239–250.

[158] Roberta Frass-Kriegl et al. "Multi-loop radio frequency coil elements for magnetic resonance imaging: theory, simulation, and experimental investigation". In: *Frontiers in Physics* (2020), p. 237.

[159] Aydan Gasimova et al. "Spatial semantic-preserving latent space learning for accelerated dwi diagnostic report generation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 333–342.

[160] Melanie Hamilton-Basich. "Hyperfine receives FDA clearance for portable MRI technology". In: *AXIS Imaging News* (2020).

[161] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[162] Salome Kazeminia et al. "GANs for medical image analysis". In: *Artificial Intelligence in Medicine* 109 (2020), p. 101938.

[163] Florian Knoll et al. "Deep-learning methods for parallel magnetic resonance imaging reconstruction: A survey of the current approaches, trends, and issues". In: *IEEE signal processing magazine* 37.1 (2020), pp. 128–140.

[164] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. "Normalizing flows: An introduction and review of current methods". In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), pp. 3964–3979.

[165] Charles Millard et al. "An approximate message passing algorithm for rapid parameter-free compressed sensing MRI". In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, pp. 91–95.

[166] M Richards. "Diagnostics: recovery and renewal". In: *Report of the Independant Review of Diagnostic Services for NHS England. NHS: Long term plan, London* (2020).

[167] Mathieu Sarracanie and Najat Salameh. "Low-field MRI: how low can we go? A fresh view on an old debate". In: *Frontiers in Physics* 8 (2020), p. 172.

[168] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502* (2020).

[169] Yang Song and Stefano Ermon. "Improved techniques for training score-based generative models". In: *Advances in neural information processing systems* 33 (2020), pp. 12438–12448.

[170] Yang Song et al. "Score-based generative modeling through stochastic differential equations". In: *arXiv preprint arXiv:2011.13456* (2020).

[171] Vera Sorin et al. "Creating artificial images for radiology applications using generative adversarial networks (GANs)–a systematic review". In: *Academic radiology* 27.8 (2020), pp. 1175–1185.

[172] Anuroop Sriram et al. "End-to-end variational networks for accelerated MRI reconstruction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 64–73.

[173] Giacomo Tarroni et al. "Large-scale Quality control of cardiac imaging in population Studies: Application to UK Biobank". In: *Scientific reports* 10.1 (2020), pp. 1–11.

[174] Maarten L Terpstra et al. "Deep learning-based image reconstruction and motion estimation from undersampled radial k-space for real-time MRI-guided radiotherapy". In: *Physics in Medicine & Biology* 65.15 (2020), p. 155015.

[175] Mark Tygert and Jure Zbontar. "Simulating single-coil MRI from the responses of multiple coils". In: *Communications in Applied Mathematics and Computational Science* 15.2 (2020), pp. 115–127.

[176] Shanshan Wang et al. "DeepcomplexMRI: Exploiting deep residual network for fast parallel MR imaging with complex convolution". In: *Magnetic Resonance Imaging* 68 (2020), pp. 136–147.

[177] Burhaneddin Yaman et al. "Self-supervised physics-based deep learning MRI reconstruction without fully-sampled data". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 921–925.

[178] Hamed Alqahtani, Manolya Kavakli-Thorne, and Gulshan Kumar. "Applications of generative adversarial networks (gans): An updated review". In: *Archives of Computational Methods in Engineering* 28.2 (2021), pp. 525–552.

[179] Laith Alzubaidi et al. "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions". In: *Journal of big Data* 8.1 (2021), pp. 1–74.

[180] Hyungjin Chung et al. "Score-based diffusion models for accelerated MRI". In: *arXiv preprint arXiv:2110.05243* (2021).

[181] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. "Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems through Stochastic Contraction". In: *arXiv preprint arXiv:2112.05146* (2021).

[182] Anonymous Reviewer CUCw. *Official Review of Paper8646 by Reviewer CUCw, Maximum Likelihood Training of Score-Based Diffusion Models*. July 2021. URL: https://openreview.net/forum?id=AkltWFnxS9&noteId=z2i6lMX-A0S.

[183] Ankan Dash, Junyi Ye, and Guiling Wang. "A review of Generative Adversarial Networks (GANs) and its applications in a wide variety of disciplines–From Medical to Remote Sensing". In: *arXiv preprint arXiv:2110.01442* (2021).

[184] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in Neural Information Processing Systems* 34 (2021).

[185] A. D. Elster. *GRE vs SE*. 2021. URL: https://mriquestions.com/gre-vs-se.html.

[186] Maryam Ghadimi and Amit Sapra. "Magnetic resonance imaging contraindications". In: *StatPearls [Internet]*. StatPearls Publishing, 2021.

[187] Graphics Media Lab Video Group. *Ways of cheating on popular objective metrics: blurring, noise, super-resolution and others*. Sept. 2021. URL: https://videoprocessing.ai/metrics/ways-of-cheating-on-popular-objective-metrics.html.

[188] Yu Guan et al. "MRI Reconstruction Using Deep Energy-Based Model". In: *arXiv preprint arXiv:2109.03237* (2021).

[189] Kerstin Hammernik et al. "Systematic evaluation of iterative deep neural networks for fast parallel MRI reconstruction with sensitivity-weighted coil combination". In: *Magnetic Resonance in Medicine* 86.4 (2021), pp. 1859–1872.

[190] Qiaoying Huang et al. "Dynamic MRI reconstruction with end-to-end motion-guided network". In: *Medical Image Analysis* 68 (2021), p. 101901.

[191] Muhammad Murtaza Khan, Bethan Pincher, and Ricardo Pacheco. "Unnecessary magnetic resonance imaging of the knee: How much is it really costing the NHS?" In: *Annals of Medicine and Surgery* 70 (2021), p. 102736.

[192] Divya Saxena and Jiannong Cao. "Generative adversarial networks (GANs) challenges, solutions, and future directions". In: *ACM Computing Surveys (CSUR)* 54.3 (2021), pp. 1–42.

[193] Yang Song et al. "Maximum likelihood training of score-based diffusion models". In: *Advances in Neural Information Processing Systems* 34 (2021).

[194] Yang Song et al. "Solving Inverse Problems in Medical Imaging with Score-Based Generative Models". In: *arXiv preprint arXiv:2111.08005* (2021).

[195] Can Tong et al. "Eigenvalue-free iterative shrinkage-thresholding algorithm for solving the linear inverse problems". In: *Inverse Problems* 37.6 (2021), p. 065013.

[196] Lilian Weng. "What are diffusion models?" In: *lilianweng.github.io* (July 2021). URL: `https://lilianweng.github.io/posts/2021-07-11-diffusion-models/`.

[197] JM Winfield et al. "Whole-body MRI: a practical guide for imaging patients with malignant bone disease". In: *Clinical Radiology* 76.10 (2021), pp. 715–727.

[198] Jinxi Xiang, Yonggui Dong, and Yunjie Yang. "FISTA-net: Learning a fast iterative shrinkage thresholding network for inverse problems in imaging". In: *IEEE Transactions on Medical Imaging* 40.5 (2021), pp. 1329–1339.

[199] Gushan Zeng et al. “A review on deep learning MRI reconstruction without fully sampled k-space”. In: *BMC Medical Imaging* 21.1 (2021), pp. 1–11.

[200] Yoshimi Anzai and Linda Moy. *Point-of-Care Low-Field-Strength MRI Is Moving Beyond the Hype*. 2022.

[201] Yutong Chen et al. “AI-based reconstruction for fast MRI—a systematic review and meta-analysis”. In: *Proceedings of the IEEE* 110.2 (2022), pp. 224–245.

[202] Edward Kuoy et al. “Point-of-Care Brain MRI: Preliminary Results from a Single-Center Retrospective Study”. In: *Radiology* (2022), p. 211721.

[203] Graciela Ramirez-Alonso et al. “Medical Report Generation through Radiology Images: An Overview.” In: *IEEE Latin America Transactions* 20.6 (2022), pp. 986–999.

[204] Mike Richards et al. “Diagnostics: a major priority for the NHS”. In: *Future Healthcare Journal* 9.2 (2022), p. 133.

[205] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[206] Maarten L Terpstra et al. “-loss: a symmetric loss function for magnetic resonance imaging reconstruction and image registration with deep learning”. In: *Medical Image Analysis* (2022), p. 102509.