Generative models for 3D Magnetic Resonance Images processing

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von

Qi Wang
aus Weinan, China

2026

Tag der mündlichen Prüfung: 08.12.2025

Dekan der Math.-Nat. Fakultät: Prof. Dr. Thilo Stehle
Dekan der Medizinischen Fakultät: Prof. Dr. Bernd Pichler

1. Berichterstatter: Prof. Dr. Klaus Scheffler

2. Berichterstatter: Prof. Dr. Thomas Wolfers

Prüfungskommission: Prof. Dr. Klaus Scheffler

PD Dr. Gabriele Lohmann

Prof. Dr. Andreas Bartels

Prof. Dr. Thomas Wolfers

Erklärung / Declaration:

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

„Generative models for 3D Magnetic Resonance Images Processing“

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

I hereby declare that I have produced the work entitled "Generative models for 3D Magnetic Resonance Images Processing", submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen, den..... .....

Datum / Date Unterschrift /Signature

Acknowledgments

I'd like to appreciate my families and friends, especially my mom Yali Zhu and my dad Dehong Wang, for their support all along my journey pursuing a higher degree and a broader vision. I'd like to extend my heartfelt thanks to my supervisors, PD Dr. Gabriele Lohmann, Prof. Dr. Klaus Scheffler, and Prof. Dr. Andreas Bartels, for their invaluable guidance and encouragement throughout the program. Additionally, I'd also like to express my gratitude to the Max Planck Institute for Biological Cybernetics, Universität Tübingen, and the utopian town of Tübingen for providing an exceptional community and sources of scientific discovery.

Abstract

With the rapid development of deep learning, image quality enhancement as a critical part of the post-processing pipeline has been revolutionized entirely by neural networks. Such surging innovation is particularly prominent in the field of computer vision. However, progress in medical imaging has been more cautious due to unique challenges posed by the nature of the data.

Especially, the use of 3D data in medical imaging significantly increases the difficulty of post-processing compared to 2D data. In medical imaging, the fidelity of the data holds more importance than the perceptual quality of images, as it directly impact the scientific and diagnostic value of the results. To better understand the underlying confounding factors and explore the potential advantages of machine learning in medical imaging post-processing, this thesis focuses on image quality enhancement tasks for structural Magnetic Resonance Imaging (MRI).

This thesis investigates three types of image quality enhancement tasks in anatomical MRI: super-resolution, retrospective motion correction, and noise removal. All these tasks can be conceptualized as inverse problems, which align well with the unsupervised learning paradigm. Traditionally, addressing these challenges has required either specialized hardware support or engineering-intensive modification to scanning protocols, demanding non-trivial technical expertise and effort. In contrast, advancements in deep learning have enabled these problems to be approached by approximating surjective mappings between input and target data distributions.

However, this approach assumes a compact representation of features in the dataset, which does not often hold in medical imaging due to variabilities such as differences in vendors or scanning protocols, can affect the data representation. This further motivates the necessity of developing data-efficient neural nets for post-processing structural MRI.

Strategically, all these tasks are approached through a three-steps framework: (1) simulating corrupted data from the ground truth, (2) training a neural network to approximate the mapping between the corrupted and the ground truth data, and (3) validating the trained model on real-world corrupted data. Despite its power and accessibility, machine learning are know to fall into the trade-off between bias and variance, which limits its generalizability to unseen data. This issue is particularly critical in medical imaging, where data scarcity is a common challenge. A successful model should therefore generalize effectively to unseen data, by learning only the most essential and representative features of the dataset.

As demonstrated by the experiments, we find that the generative models, particularly Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPMs), both exceptionally well for such inverse problems. Specifically, GANs are efficient and exhibit strong generalizability when applied to 3D data, ow-

ing to their adversarial objective function. In contrast, DDPMs, as autoregressive estimators, are better suited for 2D data due to their stable training process. Despite the demanding memory and computation overhead, improved variants scale effectively with increasing data size, making them a promising model class for integration with multi-modality data. Additionally, wavelet transform was found a powerful tool for extracting features in the frequency domain of the data. This technique serves as a plug-and-play module for generative models, further enhancing the quality of the restored images by preserving fine anatomical details, even without the need for specialized loss design.

Contents

Chapter 1

Statement of Contributions

This thesis contains both published and unpublished work. For published works, the contributions of the author are as follows:

- **Paper 1 [1]:** The author of this thesis was responsible for the conception of the research idea, experimental design, data collection, and analysis. The author also wrote the majority of the manuscript. The dataset used in the paper was from online open-source repositories, and the author ensured that all data was properly cited and acknowledged.

- **Paper 2 [2]:** The author designed the method, conducted all the experiments and evaluated the methods on various dataset. The author wrote the initial manuscript and revised it based on feedbacks from co-authors.

- **Paper 3 [3]:** The author led the project, developed the methodology, and performed all experiments. The author wrote the manuscript and revised the manuscript based on comments from co-authors. Part of the results were processed by its co-author, Julius Steiglechner, who made segmentation of the 9.4T MRI data and provided the processed data for the experiments.

1.1 Contributions of the Author

This thesis presents the original work of the author, who has claimed contribution of the co-authors in the respective papers. The writing of this thesis was done solely by the author, with close feedback and suggestions from supervisor: PD.Dr.Gabriele Lohmann. All the works presented in this thesis were conducted under the supervision of PD.Dr.Gabriele Lohmann, which are kindly supported by DFG-German Research Foundation managed by Prof.Dr.Klaus Scheffler.

The detailed contribution of the author in this thesis is as follows:

- In Chapter 6, the author conducted profound analysis of the training dynamics of GANs, which laid the theoretical foundation of the work in Paper 1 [1]. The stability analysis near equilibrium is based on local linearization of the ODE system, which is the succinct abstraction of the non-linear GAN training

dynamics. Whereas, the author acknowledge the "Dirac-GAN" [4] for formulating a linear system to simplify the parameter space of GAN and facilitate the analysis and visualization.

- In Chapter 7 and Chapter 10, the author provided a comprehensive overview of diffusion models, including their theoretical background and practical applications. The author also validates to the application of the methodology for training conditional diffusion models on simulated motion-corrupted MRI.
- In Chapter 8, the author adopted novel GAN framework consisting of three concurrently updating models to accomplish realistic super-resolution of MRI data. The author also validated the generalisability of the proposed method on unseen modalities and body-parts. To further evident the effectiveness of the third model, the author conducted TSNE visualization of the latent space of the model, which shows the generalized latent space clearly separates the different body parts and modalities. The work is published in Paper 1 [1].
- In Chapter 9, the author presented a novel approach to accomplish super-resolution and denoising at one step using GANs by modulating learned frequency in the discriminator, which is published in Paper 2 [2]. The author conducted extensive experiments both in image and frequency domain to validate the effectiveness of the proposed method.
- In Chapter 10, the author validated the naive capability of diffusion models and GANs for motion correction in simulated MRI data. Moreover, the author evaluated the trained models on real-world motion-corrupted MRI data, observing the simulated data is insufficient to represent the real-world motion patterns.

Chapter 2

Notations

Indexing

i	first order index $i \in \{1, \cdots, N\}$
j	first order index $j \in \{1, \cdots, M\}$
k	first order index $k \in \{1, \cdots, K\}$

Spaces

$\mathbb{R}$	real number space
$\mathbb{R}^+$	real number space with positive values
X	Input data space
I	Image space
$\hat{I}$	Simulated-image space
x	mini-batch of input data
$\mathbf{x}$	vector of input data
θ	parameter of neural net
ϕ	parameter of neural net
ψ	parameter of neural net
λ	eigenvalues of matrix
σ	samples from standard Gaussian distribution

functions

$f(\cdot)$	invertible function
$f(\cdot)^{-1}$	inverse function of $f(\cdot)$
$g(\cdot)$	invertible function
$\mathcal{L}$	loss function
$v(\cdot)$	vector field of neural net
$u(\cdot)$	vector field of neural net
$p(\cdot)$	probability distribution density function
$q(\cdot)$	probability distribution density function

operators

$\nabla_\phi f(\cdot)$	first order derivative of function f with respect to parameter ϕ
$\mathcal{R}$	real part of the complex number
$\mathcal{I}$	imaginary part of the complex number
$\mathbb{E}$	expectation of a function
$\int_X$	integral over sampling space X

Chapter 3

Introduction

3.1 Motivation

The evolving dynamical pace of technology has brought a lot of new opportunities and advancement in diverse scientific research topics. With the universal ability for approximating any function, deep learning has become a flexible and powerful tool in solving complex problems that were previously deemed intractable. This thesis aims to explore the application of deep learning techniques, particularly generative models, in the field of medical imaging.

As machine learning techniques are reshaping the way many traditional technique works. Unlike even 10 years ago, research in each discipline now relies more on interdisciplinary collaboration or communication than ever, especially when it comes to scientific computing and numerical solutions. Although the frontier questions remain challenging at different levels for each disciplines, it is feasible to parameterize a neural network to make predictions for solving the problem with clear objective function. Along with the research progresses, these problems become even more complex and entangled, making the traditional way of breaking them down and optimizing each one nearly impossible. However, the concept of big data has brought a more positive impact, where densely distributed data provides enriched representation with its underlying diversity. With the consensus of building big datasets for each domain, nearly all ill-posed problems can be tackled with effective results beyond traditional approaches by curating large neural networks trained on the representative dataset. With this motivation, modern deep learning methods are becoming more popular than ever in various fields, and the results are promising.

3.2 Deep learning for medical imaging

Deep learning has long been an important technology for tackling real-world scientific problems. In contrast to the early days, when deep learning models were mostly expected for proof-of-principle tasks, such as handwritten digits recognition (*e.g.* on MNIST datasets [5]), natural image classification, or instance segmentation and annotation. Similarly, in medical imaging, there has been more real-world semi-automated work on lesion segmentation, organ detection, and lesion classification.

Nowadays, deep learning tends to focus more on asking the model to learn the underlying manifold of the existing datasets, which can be helpful both in generating synthetic data samples and in understanding the distribution of the data (*e.g.* image enhancement, domain translation, visual question answering). With the promising results of recent studies, deep learning is showing its potential to bring forward post-processing steps, even taking the entire image processing pipeline to the next level. Therefore, efforts have been made to develop scalable and generalizable models, to understand data distribution, and to train cross-modality models, such as vision language models.

Outlooks. Despite the success of deep learning being applied in numerous tasks in medical imaging, research on regressive tasks has saturated the performance to the point where distinct generalisability and representative modelling are expected more than ever. Thus, the potential for development are expanded to either broaden the scope of application of AI, or to tailor the modelling specificity to particular scientific problems. A related example would be combining the learned underlying variety of attributes of MRI to extrapolate to low-field MRI for quality enhancement.

3.3 Outline

This thesis is around the topic of using different deep learning methods, especially generative models, to solve different problem in medical imaging analysis. The thesis is organized as follows:

- Chapter 4 gives a brief introduction to the background knowledge of three problem to be tacked: super-resolution, denoising, and motion correction; and fundamentals of generative models;
- Chapter 5 reviews the related works ranging from the development of generative models to the applications in medical imaging;
- Chapter 6 and 7 introduces the methods used in this thesis, including Generative Adversarial Networks (GANs) and Diffusion Models (DDPM);
- Chapter 8 presents the detailed application of GAN model in super-resolution for medical imaging;
- Chapter 9 presents the detailed application of wavelet-informed GAN model in super-resolution and denoising for medical imaging;
- Chapter 10 presents the detailed application and early investigation of both GAN and diffusion model in motion correction for simulated motion and real-world motion in medical imaging;

Chapter 4

Problem statement

This chapter contains the background and the motivation of the research questions investigated in this thesis, including learning-based method for magnetic resonance image (MRI) quality enhancement in section 4.1 and the development of effective generative models to enhance MRI quality, as in section 4.2.

4.1 Image enhancement for MRI

Image quality enhancement tasks have been extensively studied in the 2D field, but the problem becomes much more challenging when it comes to 3D MRI where practical limitations arise such as lack of data, strict law of physics, and the curse of dimensionality.

A lot of domain knowledge can be applied in the context of 3D MRI to serve as complementary information. In general, these domain techniques for image quality enhancement are non-trivial and often require additional hardware support, ranging from special design for the protocols of the signals [6] to hardware crafting and fine-tuning for RF antenna which is laborious and costly [7]. Therefore, with the increasing need for analyzing or processing high quality 3D MRI, more and more works have been done to improve the quality of 3D MRI in a post-hoc manner. Fortunately, the recent development of deep learning methods has shown better results than traditional algorithms in improving the image quality of 3D MRI, which is also the main focus of this thesis.

To shed a light on the field of combining deep learning and medical imaging, we present a set of curated generative models for 3D MR quality enhancement. In this thesis, three types of image quality enhancement are considered, they are super-resolution, motion correction, and noise removal in anatomical MRI. These tasks are promising for a wider range of applications.

These applications include image augmentation for high resolution image datasets, which is a challenge for the scarce high quality datasets and is used for training models for regression tasks, such as segmentation or classification models 4.1. In addition, super-resolution with noise removal functionality can be potentially be used to improve the quality of low-field MRI, which is more prone to noise and limited to low resolution. Moreover, correction in anatomical MRI with motion artifacts can

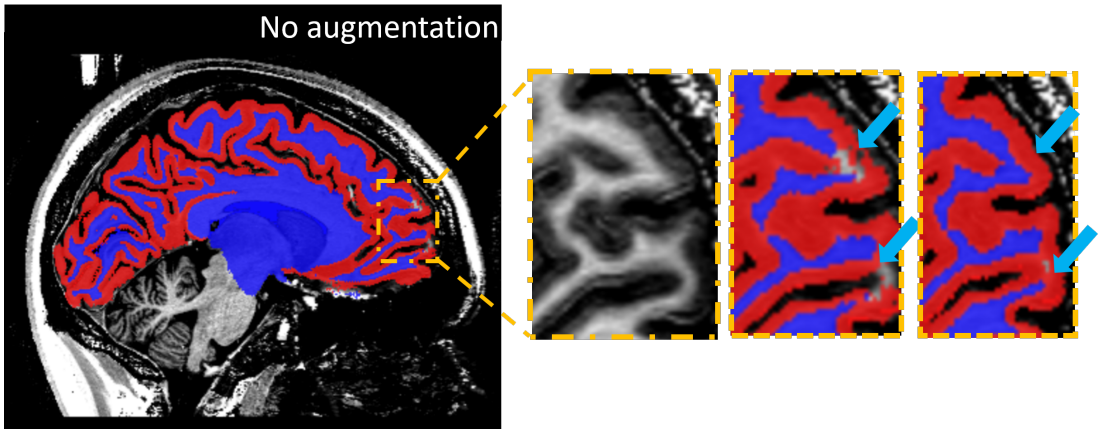


Figure 4.1: Exemplary demonstration of using the super-resolution augmentation for improving segmentation model performance in 9.4T dataset [8]. The segmentation result from the training session using augmentation (right most) shows more accurate labeling than the one without augmentation (second from the right).

help restore structural information corrupted by noise. Such restoration is essential to aid diagnosis or analysis, but has been less studied compared to motion correction in fMRI. Due to its retrospective nature and the lack of fixed reference images, the motion correction task can be more challenging than the other two tasks. Overall, we aim to demonstrate the utility of low-cost and well-generalized deep learning methods to address the aforementioned tasks in MRI image enhancement by solving inverse problems.

4.1.1 Super resolution for MRI

Super-resolution (SR) is one of the critical problems when it comes to radiology imaging, where fine anatomical details are often vital for downstream analysis. However, due to hardware limitations and physiological noise, the image quality degrades, resulting in blurred details and poorly defined boundaries.

In most of the modern SR works, simulating the low resolution image can be seen as a function $f : \mathbb{R}^d \rightarrow \mathbb{R}^l$ (here we generalise the notation $\mathbb{R}^d$ and $\mathbb{R}^l$ as image spaces for high resolution and low resolution). The solution to SR problem is formulated as approximating the inverse function $f^{-1} : \mathbb{R}^l \rightarrow \mathbb{R}^d$ of the down-sampling operator from the high-resolution ground truth $I \in \mathbb{R}^d$ to the low-resolution image $\hat{I} \in \mathbb{R}^l$ by a neural network $g_\theta(\cdot)$ (as in Eq. 4.2).

$$\hat{I} = f(I) \tag{4.1}$$

$$g_\theta(\hat{I}) = f^{-1}(\hat{I}) = I \tag{4.2}$$

The first deep learning method to tackle this problem is proposed in [9], where the effectiveness of the feed-forward convolutional neural network overcomes traditional

SR methods. Later, the application of generative adversarial network (GAN) [10] successfully shows the photorealistic SR result by exploiting the adversarial training with a GAN [11]. With the emerging popularity of the newer class of generative models, the denoise diffusion probabilistic models (DDPMs) [12] achieve comparable result to GANs [13] and show preferable training stability and better scalability, which are known to be notoriously difficult for GANs.

Unlike the rapid growth of novel methods in the 2D field, 3D models for MRI are hindered by exponentially growing parameters and the restriction for anatomical correctness. In contrast to the rapidly evolving 2D community, GANs remain a valid and effective method in the 3D MRI context, outperforming DDPMs by their significantly faster inference time. On the other hand, GANs are very flexible in terms of customization which allows the modification of their objective functions and output preferences according to specific needs.

Despite numerous work to enforce the photorealistic quality of the SR images [11, 14]. Many works enforce a detailed SR approximation by adding frequency domain constraints or operating directly on K-space information [15, 16]; some work preserves structural fidelity by constraining the gradient of images computed by a Laplace operator [17].

Regardless of the consensus on SR problem formulation, the choice of the loss function is open. Conventionally, mean square error is used to ensure the statistical similarity between the generated SR distributions and the ground truth distributions. However, it has been shown that the neural networks tend to produce blurred results due to the statistical similarity, instead of restoring the image with accurate and well-defined details. In contrast, the GAN model produces sharp image details while maintaining the statistical accuracy, but sometimes creates non-existent tissue or anatomical details in the MRI. Nowadays, researchers use a pre-trained neural network to match the similarity of the feature domain and evaluate the perceptual quality of the generated results.

4.1.2 Noise removal for MRI

Noise artefact is major cause of image quality degradation during the acquisition process in the frequency domain and has many inherent causes, including field strength, RF coils, resolutions, and gradient coils vibrations . Noisy MRI can be misleading and lead to misdiagnoses of the patients. It can also reduce the signal-to-noise ratio which smears out the visual information of certain tissues. For most of the downstream analyses, a clean visual representation of the MRI is essential for the correctness of the analysis. Therefore, noise reduction is an essential pre-processing step in many pipelines, which plays a critical role for downstream tasks, such as segmentation.

Typically, noisy MRI is a mixture of multiple noise sources, such as thermal noise from the machine, motion noise, or physiological noise from the patients. This makes the characterization of the noise type challenging, and hinders the solution to tackle specific noise type. Therefore, many well-developed, traditional, filter-based meth-

ods can mitigate the typical Rician noise [18] but fail to recover the remaining noise types in one step (as shown in the result of Chapt. 9).

To solve such a task in a learning-based method, neural networks are used to approximate a map between clean and noisy image distributions. The noise in MR image intensity is known to follow a Rician distribution, which can also be approximated by a Gaussian distribution. The noisy source distribution is easy to simulate and is most often simulated by convolving Gaussian noise with clean images. Thus, the noise removal task can be formulated as image translation between paired noisy and clean images [19, 20, 21].

Although the simulated noisy MRI synthetic data serves well for the training, the ground truth images for noisy images are difficult to acquire, making the retrospective noise cleaning task of the paired noisy and noise-free reference image insufficient. Many deep learning based methods rely heavily on the paired data from the same subject to learn a mapping between noisy and clean data distributions. In the absence of a reference-noisy image pair, it is a common practice to simulate noisy images from the clean image dataset. However, this strategy is heavily dependent on the variability scales of the noise types, and tends to overfit to certain noisy types.

Despite the success of machine learning in image processing fields, the lack of a dedicated denoising model for MRI still hampers the task. Most of the neural networks are trained to minimize the statistical discrepancies between target and source distributions, yielding a rather blurry output and ignoring the innate spatial frequency information contained in MRI. Therefore, the chapter 9 proposes a network with submodules focused on generating photorealistic MRI with noise removal functionality constrained by minimal loss of frequency information.

4.1.3 Motion correction for MRI

Motion artefact is one of the most common and important noise in anatomical MRI. It is subject to the physical motion of the subject during the long acquisition time and causes deterioration of the frequency information in the k-space, which often adds up with other types of artifacts (as is shown in Fig. 4.2). Motion in brain MRI is often considered as a random mixture of 6 rigid motion parameters (three translational direction and three rotational), neglecting the non-rigid motion caused by subtle brain pulsation. This, the motion correction, considers only the rigid body motion as in chapter 10.

Although mostly studied in the context of temporal data, the term motion correction in this thesis refers to the retrospective correction of motion artifacts in anatomical MRI not in fMRI [22]. Due to the lack of a reference image of the same acquisition, the recovery of the clean image depends heavily on the estimated retrieval of the original anatomical feature. Leveraging the universal approximation theorem of deep learning [23], it is a highly suitable method to approach the motion correction problem by approximating the clean anatomical information from the deteriorated one. In practice, it is often necessary to remove motion artifacts retrospectively to retain semantically meaningful image content for the purpose of diagnoses or further

analysis [24, 25].

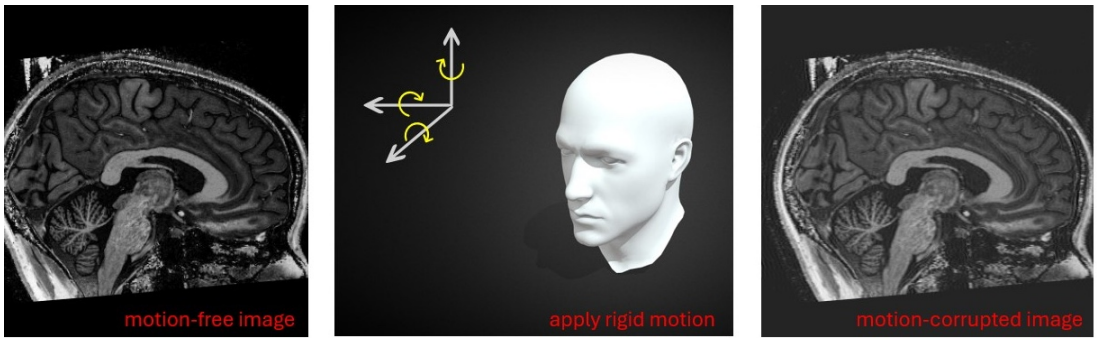


Figure 4.2: An illustration for the motion-corrupted MRI (right) due to physical motion during the acquisition (middle). The motion-corrupted image looks more blurred and has ghosting artifacts compared to the motion-free image (left).

Despite the numerous effort to restore image content free from motion-corrupted images, contaminated structural information are bounded to the fully-sampled acquisition and can vary significantly from scan to scan [26]. Similar to image denoising, motion correction suffers from not having the ground truth as a reference, complicating problem to retrieve original image information. Most of the traditional motion-correction methods are in prospective manner, where the following methods are developed [7] [27]: subjects are physically constrained to the bench to ensure a negligible head motion; measuring the real-time subtle head motion as well as breath motion by optical motion tracing and correct the motion based on the measurement of simultaneous laser tracing [28]. On the other hand, accelerated and parallel imaging methods are also introduced to minimizing the motion artifacts by shortening the acquisition time or reframing the information encoding process. These methods can mitigate the majority of the motion to certain extent, but still leaving the images suspicious to subtle motion and brings numerous amount of effort to develop the hardwares. Thereby, the main advantage of prospective over retrospective motion correction is the availability of the reference measurement, which is crucial for the latter. But retrospective motion correction is more flexible and can be applied to the existing data, which is more practical.

With emerging deep learning methods, neural networks fit well the task of retrospective motion correction for anatomical MRI by approximating the clean image from the motion-corrupted ones. The training data can be acquired by simulating the motion images by adding random motion to the reference images. The network is commonly trained to minimizing the statistical discrepancy between the reference and motion-corrupted images with perceptual constrain. However, motion-corrupted images varies in the presence of motion artifacts and severity of ghosting, which enforces the network to be generalisable to the rarely seen data. In Chapter 10, we investigated two most powerful generative models, GAN [10] and DDPM [12], for the task of motion correction in anatomical MRI.

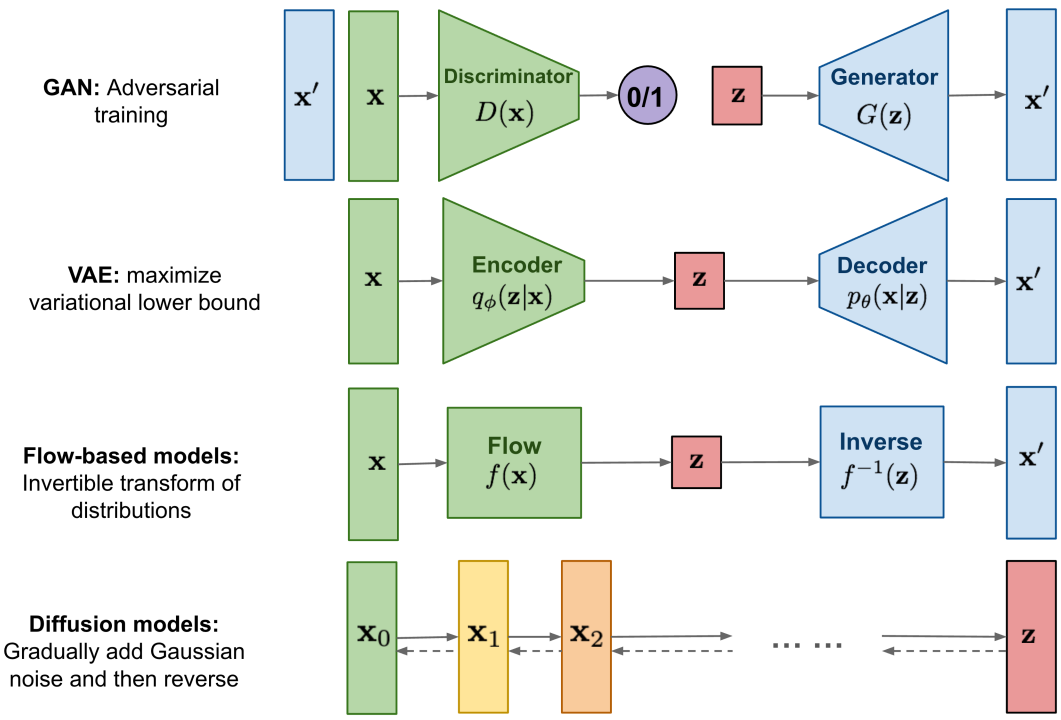


Figure 4.3: An overview of the modern generative models, including VAE, GAN, flow models, and DDPM. Image source: [32].

4.2 Deep Generative Modeling

Generative modeling is a process of approximating the underlying distribution of the dataset by minimizing the energy of an objective, of which the learned latent variables can be used to generate new samples by sampling inside it. The generative modeling can be achieved by maximizing the likelihood of the predicted data distribution, or minimizing the probabilistic discrepancies between the predicted and the true data distribution. More specifically, as are shown in Fig. 4.3, modern generative models are mainly in types of Variational Auto-encoders (VAE) [29], Generative Adversarial Networks (GAN) [10], Diffusion Probabilistic Models (DDPM) [12], Normalizing Flow [30], and Restricted Boltzmann Machines (RBM) [31]. Due to their powerfulness in capturing underlying data distribution, the main real-world usage of generative models has already expanded beyond image generation to more scientific fields, such as drug discovery, protein folding, and molecular design. Thereby, we will further introduce how the generative models are used in the context of MRI image enhancement (as in Chapter 8 9 10).

4.2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [10] are a recent class of generative models that are typically trained by two differentiable estimators, where one is responsible for generating samples and the other for distinguishing the generated samples from the true data distribution (as is shown Fig. 4.3). The GANs are firstly known for the ability of generating realistic images even at high resolution, which is difficult for previous convolutional neural nets (CNN) or VAE models. Apart from that, GANs are also naturally design to be trained for unsupervised learning as well as minimizing the entropic discrepancies between generated and original datasets, which is a significant improvement from the traditional CNN models or VAEs. Furthermore, the learned latent space of GANs is continuous and can be used for interpolation or manipulation of the generated samples, which is a significant advantage for the downstream analysis or data augmentation [33]. Despite their effectiveness, GANs are known for their notorious training instability, which is due to the "minimax" adversarial training optimization. The training process is difficult to balance between the generator and discriminator, which leads to the mode collapse or vanishing gradient problem. To mitigate this issue, many works have been proposed to stabilize the training process by enforcing additional regularization on gradients or loss functions, such as the most impactful works: Wasserstein GAN [34], spectral normalization [35], Progressive GAN [36], and BigGAN [37]. Despite the training instability, GANs are still one of the most popular generative models for data generation and image enhancement tasks, which is usually achieved by careful design of the objective functions and the network architectures. The regularization for improving the optimization process is discussed in Chapter 6 and the applications in MR image quality enhancement are shown in Chapter 9.

4.2.2 Diffusion Models

Diffusion models refer to a general class of generative models backboned by the iterative refinement of the denoising process [12, 38]. Originally inspired from non-equilibrium thermodynamics, the diffusion model is designed to learn an invertible noise schedule of the denoising process, which is then used to generate the clean image samples at inference time [39]. Each step of the denoising process is approximated by a neural network conditioned on the noisy samples from the last time step (see Fig. 4.3).

The diffusion model is known for its stable training process and the effectiveness to generate high-quality images. However, the inference time is significantly slower than GANs, which is due to the numerous amount of the iterative denoising steps [40]. From statistical modeling perspective, the diffusion models can be seen as a latent variable model, whose latent space is discretized into numbers of denoising steps and the latent states are updated sequentially during the training. Therefore, the diffusion model can be analogous to a generalization of the VAE model, whose latent space is continuous in time. Furthermore, the loss function design of the diffusion models is comparable to the one of VAE, which is optimized to maximize

the variational lower bound of the log-likelihood of the generated data distribution. Despite that, different objectives also show excellent performance, such as score-matching [38] or noise-prediction loss [12]. With that, diffusion models also benefit from the great design of VAEs, such as the manipulatable latent space and the ability to generate conditioned samples by sampling the latent space, while being able to capture the underlying data distribution by maximizing likelihood of the generated samples. Compared to GANs, diffusion might seem slow and resource-intense in training, it possess many preferable properties that enables not only scalability of the task, but also strong statistical support bounded by matured knowledge frame of statistical mechanics. Leveraging the advancement of such generative modeling approach, we investigated the applications of diffusion models in MR image quality enhancement as discussed in Chapter 10.

Chapter 5

The State-of-the-art

This chapter discusses the related works in the field of MRI image processing using generative networks as well as the development of generative models. The first half of the literature review mainly includes tasks with learning-based methods for image super-resolution, image denoising, and image motion correction. The second half focuses on the development of generative models for tackling reverse problems on MRI. The first half is divided into three sections, each focusing on one of the tasks mentioned above.

5.1 Image enhancement for MRI using Neural networks

5.1.1 Super resolution

Learning based SISR has been actively studied vastly in 2D but rarely in 3D images. In a SISR task, the neural network aims to learn a non-linear mapping from the low resolution (LR) image to its high resolution (HR) reference. This is particularly suited to the learning capability of convolution neural networks (CNNs), as demonstrated in previous studies[3,10,25,30]. In the context of 2D MRI, a squeeze-excitation attention network achieved remarkable results [29], while a transformer architecture achieved superior quality in a multi-scale network [12]. However, since downstream analysis typically requires 3D volumes of MR images, stacking 2D slices of SR results can lead to artifacts. Therefore, implementing a 3D model which operates directly on 3D volumes can naturally outperform 2D models [18].

The Chapter 8 will focus only on 3D MR image training. Most of the novel architectures have not been adapted to 3D MRI data, due to the high dimensionality and lack of data. Thus, especially for the 3D MRI domain, the GAN category remains the mainstream training method, given its efficiency and high performance. An initial work by [18] discussed the implementation of CNN models on 3D MRI data for SISR tasks, demonstrating the advanced SR results produced by neural networks over traditional interpolation methods. In [2], the authors combined Wasserstein GAN [34] training with densely connected residual blocks and reported prominent image quality. While in [26], the authors used implicit neural representation to realize SR by generating images at an arbitrary scale factor. In [22], the authors proposed to use image gradients as a knowledge prior, for a better matching to local image

patterns.

5.1.2 Noise removal for MRI

Traditional methods perform reasonably well for estimating Rician noise and its removal [18], which were mainly done by the curated non-local filter design, such as bilateral filtering [41] and median absolute deviation [42]. Some of these methods are implemented in the mainstream applications, such as adaptive non-local means denoising method in SPM12 [43].

Despite the effectiveness of the traditional methods, learning based approach are shown to outperform the classical methods, with better performance and strong flexibility to input samples. In [44], the neural net is tuned to adjust the input noise level to predict a cleaned image by a CNN with novel activation function. Multi-channel strategies combined with denoising CNN were proposed to remove Rician nosie robustly [45]. In the work of [46], the authors proposed to train a CNN to remove image with stronger Rician noise added, and their validation was performed on less noisy data and shown to be more effective when trained with less noisy data. These ideas are mostly based on the design of the training setup regarding noise-adding schemes, whereas the network model itself only consists of convolution layers and activation functions.

With the rapid development of generative models, GAN was deployed as a prevailing model for this denoising task. Under the framework of GAN, [20] introduces a cycle generative adversarial network, dubbed CycleGAN, to denoise the MRI after degraded from compress sensing. with the emerging of the class of diffusion models (DDPMs), the authors in [21] proposed to use the three-stage diffusion probabilistic model to reduce the noise in MRI, which shows a better performance and more stable training process than the GAN model.

5.1.3 MRI motion correction

Retrospective motion correction methods can be vaguely categorized into prior-information and deep-learning based approaches. As the name suggested, prior-information group leverages additional measurement of the motion parameters as a reference to correct the motion in either k-space or image space, such as navigator-based methods and optimal measure.

Conventional methods for motion correction

As a early method for tacking motion for MRI, which uses self-navigation technique that estimate motion directly from k-space trajectories, PROPELLER was proposed in [47] that acquires k-space data in concentric rectangular strips to correct through-plane motion. MOJITO [48] was proposed to detect in-plane translation using intersecting k-space trajectories and phase differences. Moreover radial 1D trajectories were proposed to replace sub-trajectories in self-navigating k-space [49, 50], which improved the variability on temporal resolution of the image. Furthermore,

[51] advanced radial trajectory scheme for both translational and rotational motion by using registration of motion data to the motion free data. Different from self-navigation, navigator echo method (NAV) requires additional measurement at the beginning or end of the actual scan, as a reference to correct the actual scan [52, 53, 54]. However, the obvious drawback of these methods is that they require additional measurement of the motion, requiring prolonged scan time. To this end, GRAPPA [55] combined navigator echoes and generalized auto-calibrating partially parallel acquisition (GRAPPA) to correct both in-plane and through-plane motion.

Optical tracking system uses cameras to track the motion of built-in marker inside MR system. Infrared tracking system was used to monitor motion [56], where two cameras were placed outside the scanner bore to detect a set of reflective markers and restrained 6 degree of freedom. As the cameras were placed inside the bore, closer to track the marker, motion tracking gained improved accuracy [57]. Furthermore, the system was further extended to a three-camera system [58] and targeted only on markers in 7T scanner. Single camera system gained more interest due to its simplicity and smaller footprint, [59] proposed to mount a single camera to head coil with checkboard calibration for motion tracking. With the development of the single camera system, [60] proposed moiré phase tracking based on patterns generated by regularly spaced elements of a transparent substrate, reducing tracking accuracy to be at 0.01mm in translation and 0.01 degree in rotation.

Provided the above-mentioned measurement, retrospective rigid-body motion correction can be done via linearly correcting the 3 translations and a phase correction in k-space [48]. This is mostly done in Cartesian sampling but can be easily generalized to other sampling schemes. However, conventional correction can only be applied for 2D in-plane translation which assumes the phase correction is fixed per k-space line or a set of them [53, 48, 50]. Rotation correction depends on the Fourier rotation theorem [61], which correspond rotation angles in image space to it in k-space.

Deep learning based motion correction

Due to the laborious process of acquiring additional measurement and costly setups for conventional motion correction, nowadays, deep learning-based methods becomes more popular in motion correction tasks. The deep learning-based methods work on two different types of data, image data and k-space data.

Formed as image-to-image tasks, modern retrospective motion correction approaches mainly utilize convolution network which provided strong capability for pattern matching and feature extraction. Image patches were sampled locally at multiple resolution of an MRI brain image, and were passed to a residual convolution network to capture the feature of motion artifacts [62]. Similarly, convolution auto-encoder was employed to extract the motion representative feature of motion artifacts and reconstruct the motion-free image by decoding the latent space [63, 64, 65, 66]. Commonly, Unet structures were utilized whose skip-connection can help to preserve detailed spatial information of the image features at different resolutions [63, 66, 67].

With the emergence of generative adversarial networks, GAN was also employed to correct the motion artifacts in MRI images [68].

5.2 Generative Models

Generative modeling is a field of machine learning that focuses on fitting the underlying distribution of the training dataset with neural nets. Currently, three classes of generative models are widely used in the field of machine learning: Energy Based Models (EBM) [69], Variational Autoencoders (VAE) [29], and Generative Adversarial Networks (GAN) [10]. Each class focus on different trade-offs between runtime, diversity, and architecture options. Being distinct in the objectives, EBM models the distribution of the dataset by minimizing the energy function, VAE learns the latent vector of the dataset by maximizing the lower bound of the likelihood, and GAN commonly simultaneously optimize on two objectives while maintaining equilibrium.

Historical generative models from statistical physics

The first generative model uses neural networks can be traced back to the Hopfield network, where the network consists of a single layer of binary-threshold "neurons" [70]. Following the development of the Hopfield network, the Boltzmann machine was proposed to learn mapping and to capture certain attributes from data, which is a stochastic version of the Hopfield network [71]. The Boltzmann machine operates in a non-deterministic way because of it layout of the neurons, which only contains visible and hidden layers, but no output layer. Moreover, minimizing the objective energy function also leads to thermal equilibrium and ensured stable optimization process. The emergence of the Boltzmann machine has laid the foundation for the development of generative models.

In modern application, however, the Boltzmann machine is computationally expensive to train, and it takes too long to train. The development of the Restricted Boltzmann Machine (RBM) [31] and the Deep Boltzmann Machine (DBM) [72] has added more layers to the network structure and made the training of generative models more efficient by adjusting the connections among layers. The modifications from RBM and DBM has enabled deep neural networks for capturing the distribution of high-dimensional dataset (such as MNIST and NORB). These class of model optimizing the energy function were named as Energy Based Model (EBM) [69] ever since.

Modern development of generative models

Unlike the early energy-based models, some generative models focusing more on learning the latent vector of the dataset which enables the sampling from latent representation, such as the Variational Autoencoder (VAE) [29]. VAE possess the advantage of explicitly allowing to sample in the latent space and decode samples with

respective features. By leveraging the variational inference, β-VAE [73] was proposed to control the disentanglement of the latent space, which allows the model to learn the independent factors of variation in the dataset. Masked Autoencoder (MAE) [74] was proposed to learn the latent representation of the dataset by masking partly the input to the autoencoder, which allows the model to learn more generalizable latent representation of the dataset while throughputting less information. This is particularly useful when it comes to the pretraining of larger models, where a useful latent representation along with reasonable computation cost is required.

Based on the game theory, the Generative Adversarial Networks (GAN) have been proposed, where a system of two networks, a generator and a discriminator, are competing against each other to generate realistic samples [10] from source distribution (*e.g.* standard Gaussian) [10]. Compared to VAE and early EBMs, GAN has improved the performance of the generative modeling by a huge margin, and has been widely adopted in various scientific disciplines, such as image generation [75], image super-resolution [11], and image-to-image translation [76]. A more recently variation of generative models is the combination of the concept of VAE and Random Markov Field, which is denoising diffusion probabilistic models (DDPM) [12]. DDPM is a generative model that can generate samples by iteratively applying trained network and recover the source image from the target ones.

Chapter 6

Generative Adversarial Nets

6.1 Introduction

To better understand the reason behind the design choice of our model and their detailed implementation, this chapter provides a comprehensive review of the stability and convergence property of the Generative Adversarial Networks (GANs) [10]. During the porpular time of GAN, many variants were proposed to stabilize the training process, such as the Wasserstein GAN (WGAN) [34], the instance noise, and the non-saturated GAN. Heuristically, WGAN was the most popular method for training a converged and balanced GAN. However, we noticed the qualitative results using WGAN yields many problems, such as mode-collapse and poor image quality. We suspected the reason behind this is the instability of the GAN training. In this chapter, we proved the instability of WGAN and convergence of the instance noise, using simplistic setup as Dirac-GAN [4].

6.2 Non-Convex optimization

Non-convex optimization aims to optimize the function that is not convex, which is not restricted to concave nor convex functions (exemplary samples of convex and non-convex function landscape is shown in Fig. 6.1). It is an open question to find out the solution to such a problem, due to its nature of non-guaranteed convergence to global minima. Aside from the non-convergence, non-convex functions are also known to be unstable, as the optimization process can be trapped in the saddle points, where the gradient of the function is zero but not the global minima.

Despite the difficulties, many methods were proposed to improve the optimization process. In the context of deep learning, gradient descent is the most effective way for non-convex optimization. The optimization for GAN's objective is typically non-convex, the mini-max optimization.

6.3 Adversarial Training

Adversarial training is the core mechanism in the GAN system. Often, the discriminator f_ψ is used to evaluate the generated samples from the generator g_ϕ, with

ψ and ϕ as the models' parameters respectively, while the generator is crafted to generate samples that are close to the training samples.

The adversarial game is formulated as a mini-max optimization, as the objective function Eq. 6.1 was proposed in the original GAN paper [10]. During the generation process, random samples $\tilde{x}$ are drawn from a prior distribution $p_{\tilde{x}}(\tilde{x})$(*e.g.* $\tilde{x} \sim \mathcal{N}(0,1)$) and are fed to the generator. The loss function penalizes the discriminator f_ψ from misclassifying samples of the real data distribution p_{data} to be fake, and generated samples $g(\tilde{x})$ to be true, by finding the optimizing f_ψ to maximize the cross-entropy. On the other hand, the loss function encourages the generator g_ϕ to generate samples that are prone to be classified as true samples by the discriminator. The loss function is expressed as follows[10].

$$\min_g \max_f \mathcal{L}(f,g) = \mathbb{E}_{x\sim p_{data}} \log[f_\psi(x)] + \mathbb{E}_{\tilde{x}\sim p_{\tilde{x}}(\tilde{x})} \log[1 - f_\psi(g_\phi(\tilde{x}))] \tag{6.1}$$

The above objective function is expected to converge to an equilibrium point, where the parameters ϕ, ψ are at their optimum in the local area and stop to update. As a adversarial nature of the mini-max game, the objective of the generator g_ϕ and the discriminator f_ψ can be simplified as zero-sum game: $f = -g$. Thus, the Nash-equilibrium is a point $\bar{x} = (\bar{\phi}, \bar{\psi})$ that satisfies the condition that, ϕ is the local maxima given the optimal $\bar{\psi}$, vice versa for ψ. This can be formulated as the following:

$$\bar{\phi} \in \operatorname*{argmin}_\phi g(\phi, \bar{\psi}) \quad \text{and} \quad \bar{\psi} \in \operatorname*{argmax}_\psi f(\bar{\phi}, \psi). \tag{6.2}$$

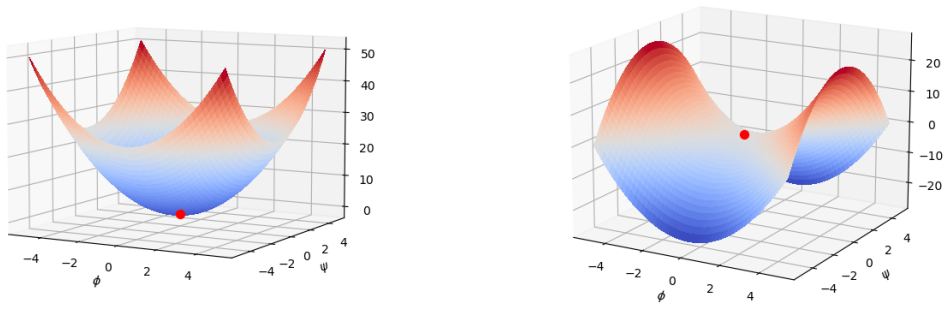
As is described by Eq. 6.2, a generator is optimized to minimize the adversarial loss when the discriminator reaches optimal value, while a discriminator is optimized to maximize the adversarial loss when the generator converge to the optima. During the adversarial optimization process of converging to saddle point, GAN might face common problems that cause the process prone to diverge. The most common problems are gradient vanishing /exploding, deviation from local minima, or oscillating around saddle point but never converge. All these common problems can be tackled by the stability analysis of the GAN training process in the following sections.

6.4 Convergence of the GAN training

As GANs are known for its unstable training dynamics, the convergence is thus not guaranteed. Many studies of the GAN stability show prevailing results, amongst those we compared several simple yet effective methods against the most popular one for detailed analysis, the Wasserstein GAN. In detail, we compared the weight-regularized, instance-noise, non-saturated, and Wasserstein GAN variations.

6.4.1 GAN as a fixed point system

Formulated in game theory, optimization on GAN objective can be seen as a result of fixed point algorithm. In this process, the function of the dynamic system



(a) (b)

Figure 6.1: Exemplary loss landscape of the convex and saddle point optimization on the parameter space. The optimal points for both functions are at the origin (marked in red). Note that the saddle point is not minimum nor maximum, but a point where the gradient is zero and is Nash-equilibrium.

refers to the entire adversarial process and the equilibrium state as the point with 0-value for the velocity, namely the discriminator and the generator can not further improve their utility unilaterally. As a theoretical solution to their convergence, their time-derivatives for both player should become zero. Due to the minimax nature of the GAN objetive function, the gradient of each player has the opposite sign to the other player. Therefore, its *vector field* can be formulated as follows:

$$v(\phi, \psi) := \begin{pmatrix} -\nabla_\phi \mathcal{L}(\phi, \psi) \\ \nabla_\psi \mathcal{L}(\phi, \psi) \end{pmatrix} \tag{6.3}$$

Simultaneous gradient descent is applied to optimize the GAN system, in the form of $(\phi, \psi)^{i+1} = (\phi, \psi)^i + \eta v(\phi, \psi)$, for $i = 0, 1, 2, \cdots, N$ and η as learning rate.

As a local linearization of the non-linear system, the following analysis is based on the local linearization around the fixed point at $(0, 0)$ of the GAN system. The Nash-equilibrium point of the GAN system is formulated as a fixed point[1] (ϕ^*, ψ^*) of the system, whose convergence heavily depends on the sign of the real-part of the Jacobian matrix for the vector field $v'(\phi, \psi)$.

6.4.2 Stability for continuous time system

The optimization process of the GAN system can be simplified as a continuous-time system. This helps to understand the GAN optimization problem through the lense of characterization of ODE.

Consider Eq. 6.3 as an linear ODE, the current states of the parameters $\mathbf{x}$, their

[1] at the fixed point, the status of the system does not change over time anymore, *i.e.* $v(\phi^*, \psi^*) = f(\phi, \psi) = 0$, with initial condition $f(0, 0) = (\phi^*, \psi^*)$, $f(t, t) = (\phi^*, \psi^*)$ for all time t.

first-order derivatives of time, and the updating rule **A**, expressed in the following ordinary differential equation(ODE):

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{6.4}$$

with **A** being iteratively updating coefficient matrix, and $\mathbf{x} = \Big(\phi, \psi\Big)^T$ corresponds to the status of the parameters in the GAN networks.

Numerically, the continuous system converges as a balanced system if, for all the eigenvalues of the Jacobian matrix of A, their real-parts have negative values; it diverges if all real-parts have positive values; and it can be both cases if zero values appear in the real-parts. With the origin point being the fixed point of the system, the integral curves (or trajectories) of the system states either converge or oscillate under the characterization of the vector fields. Here four most common dynamics are illustrated in Fig. 6.2, where a sink system converges to the fixed point with the real parts of the eigenvalues being negative (Fig. 6.2(a)), another sink system converges with circular trajectory due to its non-zero imaginary part of eigenvalues (Fig. 6.2(d)), a saddle point system has both convergent and divergent directions with opposite signs of the real parts (Fig. 6.2(b)), and a centered system has a circular trajectory around the fixed point with zero real parts (Fig. 6.2(c)).

This concept is widely adopted in many GAN variations, which decides whether the integral curve of the gradient vector field will converge to the Nash-equilibrium or not.

6.4.3 Stability for discrete time system

Despite the overall characterization for GAN convergence by its vector field, more detailed numerical analysis is needed for GAN converging behavior. In the actual implementation, where the gradient descent and weights updating are done via discrete steps. Similar to the stability analysis in discrete-time system, the GAN optimization approaching Nash-equilibrium is asymptotically stable if the eigenvalues of the Jacobian matrix of the vector field $v^{'}(\phi.\psi)$ are both all negative and are within the unit circle (*i.e.* $|\lambda| \leq 1$).

6.4.4 Toy example GAN dynamics–"Dirac-GAN"

The Dirac-GAN [4] was proposed to illustrate the problem of GAN dynamics by defining the generator and discriminator as one-parameter models. Specifically, the output distribution from the generator is defined as the parameter ϕ of the generator, such that the probability of ϕ being sampled is 1; and a linear discriminator with parameter ψ defined as the function: $f_\psi(x) = \psi \cdot x$, and real data is sampled from the distribution function: $p_x = \delta_0$ centered at 0. The Eq. 6.1 can be simplified as follows:

$$\mathcal{L}(\phi, \psi) = h(\psi \cdot \phi) + h(0) \tag{6.5}$$

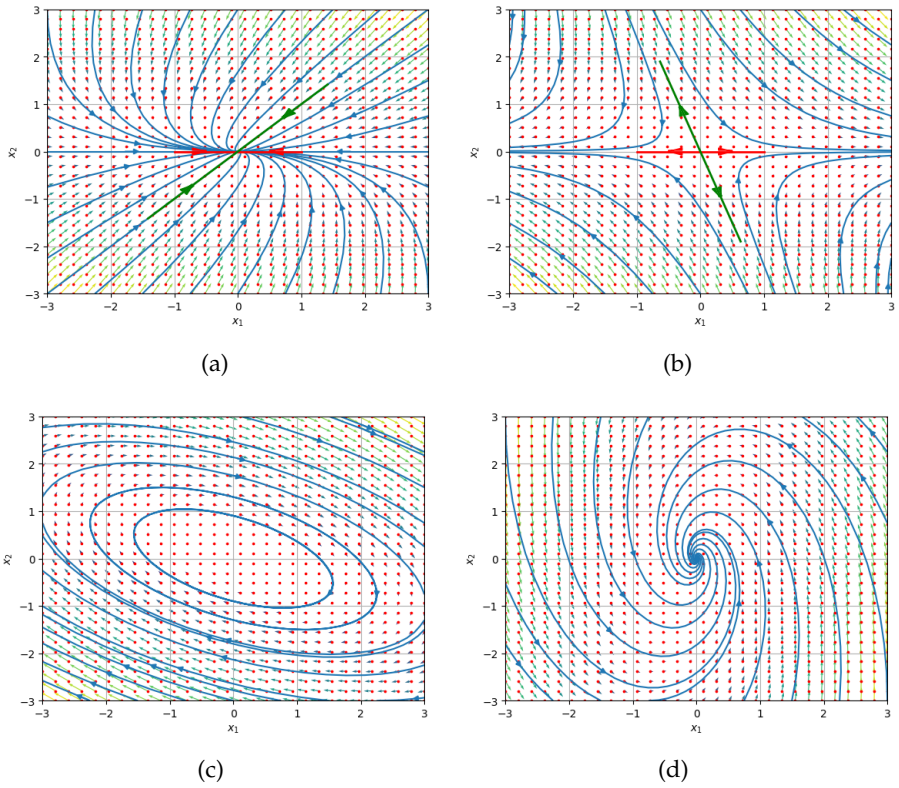


Figure 6.2: Different cases of vector field convergence to fix point. Top left: convergent system, top right: divergent system, bottom left: divergent system, and bottom right: a saddle point system. (Note: the eigenvectors are shown in the subfigures at the top row, indicating the converged direction of the vector field.)

where $h(\cdot)$ is a real-value function, such as $\log(\cdot)$ in Eq. 6.1. In the followings we use the sigmoid function: $h(t) = -\log(1+\exp(-t))$ as was in the original GAN paper [10]. Its vector field and the corresponding Jacobian can be written as

$$v(\phi,\psi) = \begin{bmatrix} -\psi h^{'}(\phi,\psi) \\ \phi h^{'}(\phi,\psi) \end{bmatrix}$$

and

$$v^{'}(\phi,\psi) = \begin{bmatrix} -h^{''}(\phi\psi)\psi^2 & -h^{'}(\phi\psi) - h^{''}(\phi\psi)\phi\psi \\ h^{'}(\phi\psi) + h^{''}(\phi\psi)\phi\psi & h^{''}(\phi\psi)\phi^2 \end{bmatrix}$$

As was defined in the objective function (Eq. 6.5) $\mathcal{L}(0,\phi) = \mathcal{L}(\phi,0) = h(0) = const, \forall\phi,\psi \in \mathbb{R}$. Therefore, the equilibrium point is $(\phi^*,\psi^*) = (0,0)$, and the Jacobian matrix of the vector field in Eq. 6.3:

$$v^{'}(\phi,\psi)|_{\phi=0,\psi=0} = \begin{pmatrix} 0 & -h^{'}(0,0) \\ h^{'}(0,0) & 0 \end{pmatrix} \tag{6.6}$$

has two eigenvalues at this point $\pm h^{'}(0,0)i$. Therefore, the integral curves of the gradient of the vector field do not converge to Nash-equilibrium. Instead, it forms a centered field which oscillates around the optimum but never converge.

6.4.5 Stability for GAN optimization

The optimization strategies for GAN include mainly two strategies: simultaneous update and alternating update. The simultaneous update is to update the generator and the discriminator at the same time, whereas the alternating update is to update the generator per multiple updates of the discriminator. Empirically, the alternating update is more stable than the simultaneous update, as the discriminator is updated more frequently than the generator and learns better to distinguish between generated and original distributions. However, the alternating update is more computationally expensive than the simultaneous update and requires more careful configuration for the updating ratio between the generator and the discriminator. These factors cause the alternating update strategy much harder to use that the simultaneous update.

In this chapter, we only focus on the simultaneous update, as it is the most common way to train GANs. The simultaneous update is formulated by an updating function $F_h(\phi,\psi)$ as follows, with η being the step size:

$$F_h(\phi,\psi) = \begin{pmatrix} \phi - \eta\nabla_\phi\mathcal{L}(\phi,\psi) \\ \psi + \eta\nabla_\psi\mathcal{L}(\phi,\psi) \end{pmatrix} \tag{6.7}$$

Substituting in the vector field $v(\phi,\psi)$ from Eq. 6.3, the updating function can be rewritten as

$$F_\eta(\phi, \psi) = (\phi, \psi) + \eta v(\phi, \psi). \tag{6.8}$$

With its gradient being $F_\eta^{'}(\phi^*, \psi^*) = I + \eta v^{'}(\phi^*, \psi^*)$, the eigenvalues of the Jacobian of the update function $F_\eta(\phi, \psi)$ is given by $\lambda = 1 + \eta\mu$ with μ the eigenvalues of the Jacobian of the velocity field of the training system $v^{'}(\phi^*, \psi^*)$ from Eq. 6.6. With the stability assumption in Section 6.4.3, substitute μ with $|\lambda| < 1$, the eigenvalues of the Jacobian of the update function $F_\eta(\phi, \psi)$ are within the unit circle, and the system converges to the Nash-equilibrium [4].

$$\eta < \frac{1}{|\Re(\lambda)|} \frac{2}{1 + (\frac{\Im(\lambda)}{\Re(\lambda)})^2} \tag{6.9}$$

6.5 Instability of the Wasserstein GANs

In the GAN case, the absolute value of the eigenvalues of the updating function $F_\eta(\cdot, \cdot)$ evaluated at the Nash-equilibrium point can be rewritten as $\sqrt{1 + \eta^2 h^{'}(0,0)^2}$. According to the Lyapunov stability theory, the states space model (Eq. 6.4) of the updating function is only asymptotically stable when all eigenvalues of the matrix A: have negative real parts, and their modulus are smaller than one. Therefore, the updating function is evidently unstable, due to its eigenvalues' modulus are not smaller than one, independent of the learning rate η.

Specifically, one popular GAN variant, Wasserstein GAN, was proposed to use critic network without non-linear activation function for the output. The recursive updating process of the Wasserstein GAN in Dirac-GAN setup can be expressed as:

$$\begin{pmatrix} \phi_{k+1} \\ \psi_{k+1} \end{pmatrix} = \begin{pmatrix} 1 & -\eta \\ \eta & 1 \end{pmatrix} \begin{pmatrix} \phi_k \\ \psi_k \end{pmatrix} \tag{6.10}$$

Recall that, at the equilibrium point, $(\phi_k, \psi_k) = (0, 0)$. Assume the discriminator to be converged, $\lim_{k\to\infty} \psi_k = 0$, there exists a point k_0 such that $\forall k \geq k_0$, $|\psi_k| < 1$. The updating function can be rewritten as

$$\begin{pmatrix} \phi_k \\ \psi_k \end{pmatrix} = A^{k-k_0} \begin{pmatrix} \phi_{k_0} \\ \psi_{k_0} \end{pmatrix} \text{ with } A^{k-k_0} = A = \begin{pmatrix} 1 & -\eta \\ \eta & 1 \end{pmatrix} \tag{6.11}$$

The absolute value of the eigenvalues of A are given by $|\lambda| = \sqrt{1 + h^2} > 1$. This indicates the divergence of the Wasserstein GAN training.

6.6 Instance noise for stability

When using instance noise sampled from Gaussian distribution with zero mean and variance σ^2. The objective of the Dirac-GAN with instance noise can be formulated as

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2)}h(\epsilon\psi) + \mathbb{E}_{\tilde{x}\sim\mathcal{N}(0,\sigma^2)}h(-\tilde{x}\psi) \tag{6.12}$$

with Gaussian noise added to both the real data and the generated data($\tilde{x}, \epsilon \sim \mathcal{N}(0,\sigma^2)$), $h(\cdot)$ as a sigmoid activation function.

The corresponding vector field is given by

$$\tilde{v}(\phi,\psi) = \mathbb{E}_{\epsilon,\tilde{x}}\begin{pmatrix} -\psi h^{'}(\epsilon\psi) \\ \epsilon h^{'}(\epsilon\psi) - \tilde{x}h^{'}(-\tilde{x}\psi) \end{pmatrix}. \tag{6.13}$$

Then the Jacobian of it becomes

$$\mathbb{E}_{\epsilon,\tilde{x}}\begin{pmatrix} -h^{''}(\epsilon\psi)\psi^2 & -h^{'}(\epsilon\psi) - h^{''}(\epsilon\psi)\epsilon\psi \\ h^{'}(\epsilon\psi) + h^{''}(\epsilon\psi)\epsilon\psi & \epsilon^2 h^{''}(\epsilon\psi) + \tilde{x}^2 h(-\tilde{x}\psi) \end{pmatrix}. \tag{6.14}$$

Therefore, the eigenvalues of the Jacobian of the vector field $\tilde{v}(\phi,\psi)$ evaluated at $\phi = \psi = 0$ is given by

$$\lambda_{1,2} = h^{''}(0)\sigma^2 \pm \sqrt{f^{''}(0)^2\sigma^4 - h^{'}(0)^2}. \tag{6.15}$$

Given that, all eigenvalues of the Jacobian of the vector field $\tilde{v}(\phi,\psi)$ have negative real parts at the equilibrium point, if $h^{''}(0) < 0$ and $\sigma > 0$, which holds true for aforementioned $h(\cdot)$. Hence, the instance noise can stabilize the GAN training and ensures the locally convergent for small enough learning rate.

6.7 conclusion

This chapter provides a comprehensive overview of the stability and convergence property of the GANs. We theoretically proved the instability of the Wasserstein GAN and the convergence of the instance-noise trick using the Dirac-GAN setup for simplicity. The stability analysis of the GAN training process is crucial for understanding the convergence of the GAN training process. Instance noise is a simple yet effective method to stabilize the GAN training process, yielding compelling results. The stability analysis of the GAN training process can be further extended to more complex GAN models, where deeper architecture are prone to cause gradient vanishing. The stability analysis of the GAN training process is an essential step towards understanding the convergence of the GAN training process to facilitate improving the performance of the GAN models.

Chapter 7

Denoising Diffusion Probabilistic Models(DDPM)

7.1 Introduction

As an emerging class of generative model, denoised diffusion probabilistic models (DDPMs) show predominant generation quality and scalability while being stable and easy to train. This unleashes the wide utility of the DDPM over GAN models for the image generation and image translation tasks in almost every field of application. However, the theoretical support for DDPM are not trivial to understand but the detailed design of the process might influence the models' performance. Thus, going through the theoretical background helps to appreciate the motivation and the design choices of the DDPM model in real-world problem, such as retrospective motion correction in MRI.

This chapter aims to provide a comprehensive understanding of math behind the DDPM model, including the forward and reverse diffusion process, and the design choices of the loss function. The forward process aims to add annealed Gaussian noise to consecutively bridge the original image to a standard Gaussian distribution, while the reverse process aims to approximate the original image from the noisy sample by a learned neural network 7.1. The loss function is designed to minimize the negative log-likelihood of the predicted original sample, by minimizing the evidence lower bound of the distribution.

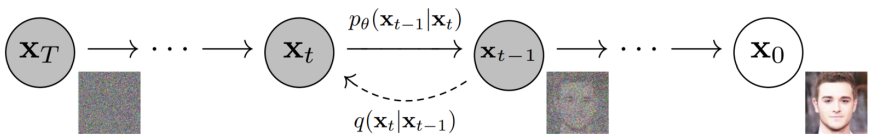


Figure 7.1: The diagram of the DDPM model. The forward process adds annealed Gaussian noise to the original image till it becomes pure Gaussian (from X_0 to X_T), while the reverse process approximates the original image from the noisy sample (from X_T to X_0). Image source: [12]

7.2 Noise simulation: The forward diffusion process

The diffusion process of the model aims to add annealed Gaussian noise at discretized time steps, such that the original images (or their probablity distribution) eventually becomes pure noise (or a standard Gaussian distribution). Consider individual group of samples at the time steps: $X_0, X_1, \cdots, X_{t-1}, X_t, \cdots, X_T$, where X_0 are the data samples and X_T are samples of pure Gaussian noise. The noise-adding process is scheduled with time-dependent variance $\{\beta_t \in [0,1]\}_{t=1}^T$ for each transition:

$$q(x_t|x_{t-1}) = \mathcal{K}(x_t|x_{t-1};\beta_t) \tag{7.1}$$

where $\mathcal{K}$ denotes a Markovian kernel for the process.

In the design of DDPM model, Gaussian transition kernel is used, *i.e.* $\mathcal{K}(x_t) = \mathcal{N}(x_t;\tilde{\mu}_t,\Sigma_t)$ with $\tilde{\mu}_t$ and Σ_t for its time-dependent mean and variance. Thus, the sequential multiplication of probability density from each noise-adding steps can be seen as the probability of a group of all $x_t, t \in [1,T]$ conditioned on the original images x_0, expressed as:

$$q(x_1,\cdots,x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{7.2}$$

As the mean and variance matrix of the isotropic Gaussian noise distribution being parameterized as[1]: $\tilde{\mu_t} = \sqrt{1-\beta_t}x_{t-1}$ and $\Sigma_t = \beta_t I$, the transition probability can be rewritten as in Eq. 7.3. The noise schedule are design to be close to 0 at time $t = 0$, and linearly growing to a fixed value at $t = T$. Namely, at the beginning the mean of the Gaussian kernel is close to 1, so as to not shifting the data distribution too much; the variance is close to 0, meaning the noise for the start of the process is very small.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t;\sqrt{1-\beta_t}x_{t-1},\beta_t I) \tag{7.3}$$

With the transition probability, the predefined Gaussian probability path, and the original data distribution, the marginal probability of the forward process at any time step t can be described as Eq. 7.4. To simplify the notions used, by defining $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$, the marginal probability can be expressed as:

$$\begin{aligned} q(x_t|x_0) &= \mathcal{N}(x_t;\prod_{i=1}^{t}\sqrt{\alpha_t}x_0,(1-\prod_{i=1}^{t}\alpha_t)I) \\ &= \mathcal{N}(x_t;\sqrt{\bar{\alpha}_t}x_0,(1-\bar{\alpha}_t)I) \end{aligned} \tag{7.4}$$

To this end, we already have a forward simulation of the noise-adding process, which is conditioned on the original data distribution $q(x_0)$.

[1]Scaling factor $\sqrt{1-\beta_t}$ is used to retain numerical stability from exploding variance.

7.3 Image generation: Reverse diffusion process

Even though the forward process is almost deterministic, the reverse process is not as convenient to compute. Formed as a Markovian chain, it is possible to revert the transition kernel to achieve the reverse process, the time-reversal process for the forward process. By Bayes' rule, the reverse process can be expressed as the posterior probability of forward probability, conditioned on the original data distribution, which is not expressed in closed form in real-world cases. Thanks to the learning ability of the neural network, we use the model to approximate the original data samples to formulate a predicted reverse process.

According to the Bayes' rule, the reverse process heavily depends on the marginal likelihood at adjacent time steps, which is only achievable by integrating the conditional distributions $q(\cdot|x_0)$ over the entire distribution of samples X_0 as in Eq. 7.5. Note, the conditional probabilities $q(\cdot|x_0)$ are the same as forward process, which is known from Eq. 7.4. Thus, the reverse transition probabilistic kernel can be expressed without the dependency on the probability of the former time step, but only on the original data distribution $q(x_0)$.

$$\begin{aligned} q(x_{t-1}|x_t) &= \frac{q(x_t|x_{t-1}) \cdot q(x_{t-1})}{q(x_t)} \\ &= \frac{q(x_t|x_{t-1}) \cdot \int_{X_0} q(x_{t-1}|x_0) q(x_0)\, \mathrm{d}x_0}{\int_{X_0} q(x_t|x_0) \cdot q(x_0)\, \mathrm{d}x_0} \end{aligned} \tag{7.5}$$

In practice, the data distribution $q(x_0)$ is not easy to be formulated or integrated with, *e.g. not like any known distribution with formulated expression*. Therefore, the reverse transition $q(x_{t-1}|x_t)$ is hard to compute due to a lack of expression for original data distribution. Whereas, this shortcoming can be tackled by designing another process which closely resembles the reverse process. Except for that the original data distribution is approximated (noted as $\hat{x}_\theta$) by a neural network (with parameters θ) and, thus, the approximated denoising process $p_\theta(x_{t-1}|x_t)$ can be calculated from noisy steps to the data distribution $p_\theta(x_0)$ (Eq. 7.6 and Eq. 7.7).

The approximated transition Markovian kernel can be expressed as Eq. 7.6. Likewise, the reverse Markovian chain can be rewritten as joint probability of the reverse process as in Eq. 7.7. Here, the prior distribution $p(x_T)$ is assumed to be the same as $q(x_T)$ when infinitesimal time steps are considered.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{7.6}$$

$$p_\theta(x_0 \cdots x_T) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{7.7}$$

Known from the forward process, the reverse process of noise-adding process,

which is conditioned on the input sample x_0, can be written as Eq. 7.8.

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_{t-1}|q(x_0))q(x_t|x_{t-1}, x_0)}{q(x_t|x_0)} \tag{7.8}$$

the individual term can be expressed as:

$$q(x_{t-1}|q(x_0)) = \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0, (1 - \bar{x}_{t-1})I)$$
$$q(x_t|x_0) = \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{x}_t)I)$$

since $q(x_{1:T}|x_0)$ is Markovian, the conditioned probability $q(x_t|x_{t-1}, x_0) = q(x_t|x_{t-1})$ can be expressed as the following:

$$q(x_t|x_{t-1}, x_0) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \tag{7.9}$$

With all the above terms, the reverse process can be expressed as Eq. 7.10.

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t I) \tag{7.10}$$

With the empirical mean and variance terms:

$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t x_0 + \sqrt{1 - \beta_t}(1 - \bar{\alpha}_{t-1})x_t}{1 - \bar{\alpha}_t} \tag{7.11}$$

$$\tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t \tag{7.12}$$

To approximate the drift term that dominant the reverse process. Substituting real data sample x_0 with predicted sample $\hat{x}_\theta(x_t, t)$:

$$\mu_p(x_t, t, \theta) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t \hat{x}_\theta(x_t, t) + \sqrt{1 - \beta_t}(1 - \bar{\alpha}_{t-1})x_t}{1 - \bar{\alpha}_t} \tag{7.13}$$

We got the expression for the mean term μ_p of the learned reverse process p_θ. Similarly, the neural network is used for approximating the reverse process, where the mean $\mu_p(x_t, t, \theta)$ is learned by the neural net θ, simplified as μ_θ. Thus, the learned reverse process can be formulated as in Eq. 7.14, except for the noise inside the network's drift need to be learned $\mu_\theta(x_t, t)$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \tilde{\beta}_t I) \tag{7.14}$$

Similar to the idea of stochastic process, the reverse process of diffusion can be expressed in Langevin equation [77]. The mathematical framework used for formulating the diffusion model is scaffolded by a time-reversal version of Langevin dynamics. As the noisy sample can be expressed by iteratively adding standard Gaussian noise to the original sample (as in Eq. 7.15), where the noise ϵ_t comes from standard Gaussian with mean and variance following the predefined schedule. The original

sample x_0 can thus be expressed by the noisy sample at any time point x_t with known noise coefficient $\bar{\alpha}_t$ and simulated noise ϵ_t[1] (as in Eq. 7.16). Similarly, the predicted sample $\hat{x}_\theta$ also can be expressed by a predicted noise step $\hat{\epsilon}_\theta$ with the same noise coefficient $\bar{\alpha}_t$ (as in Eq. 7.17).

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t,\ \epsilon_t \sim \mathcal{N}(0, I) \tag{7.15}$$

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}x_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\epsilon_t \tag{7.16}$$

$$\hat{x}_\theta = \frac{1}{\sqrt{\bar{\alpha}_t}}x_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\hat{\epsilon}_\theta \tag{7.17}$$

The empirical and the learned mean of the Gaussians should be written as the following:

$$\begin{aligned} \tilde{\mu}(x_t, x_0) &= \frac{1}{\sqrt{\alpha}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t) \\ \mu_\theta(x_t, t) &= \frac{1}{\sqrt{\alpha}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_\theta(x_t, t)) \end{aligned} \tag{7.18}$$

Therefore, the reverse denoising sampling (as in Eq. 7.14) can be analogous to Langevin process and be described as the following:

$$x_{t-1} = \frac{1}{\sqrt{\alpha}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_\theta(x_t, t)) + \sqrt{\tilde{\beta}_t}z,\ z \sim \mathcal{N}(0, I) \tag{7.19}$$

With the learned iterative noise $\hat{\epsilon}_\theta$, we can express the reverse process iteratively from samples at anytime point x_t to the original sample x_0, as in Eq. 7.19. Eventually, we noticed the key component of the reverse process is the learned noise $\hat{\epsilon}_\theta$, which is the key to approximate the original sample from the noisy sample.

7.4 Design choices of the loss

The objective function of the training process is to minimize the negative log-likelihood of the distribution, by minimizing its evidence lower bound, of the predicted original sample $p_\theta(x_0)$. According to the Markovian property of the diffusion process 7.7:

$$p_\theta(x_0) = \int p_\theta(x_{0:T})dx_{1:T} = \int_{X_1} \cdots \int_{X_T} p_\theta(x_0 \cdots x_T)dx_T \cdots dx_1 \tag{7.20}$$

Expanding the log-likelihood of the approximated original data distribution (abbreviated as $\log p(x)$) by Eq. 7.20:

[1] note that the noise scheduling only controls the statistical properties of the Gaussian distribution, whereas the specific noise samples are intractable and that's why ϵ_t is introduced.

$$
\begin{aligned}
\log p(x) &= \log \int p(x_{0:T}) dx_{1:T} \\
&= \log \int \frac{p(x_{0:T}) q(x_{1:T}|x_0)}{q(x_{1:T}|x_0)} dx_{1:T} \\
&= \log \mathbb{E}_{q(x_{1:T}|x_0)} \Big[\frac{p(x_{0:T})}{q(x_{1:T}|x_0)} \Big] \\
&\geq \mathbb{E}_{q(x_{1:T}|x_0)} \Big[\log \frac{p(x_{0:T})}{q(x_{1:T}|x_0)} \Big] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)} \Big[\log \frac{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})} \Big] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)} \Big[\log \frac{p(x_T)}{q_{x_T|x_0}} + \sum_{t \geq 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q_{x_{t-1}|x_t, x_0}} + \log p_\theta(x_0|x_1) \Big] \\
&= -\mathbb{E}_{q(x_{1:T}|x_0)} \Big[\underbrace{\mathcal{D}_{KL}(q_{x_T|x_0}||p_\theta(x_T))}_{\mathcal{L}_{\text{prior match}}} + \underbrace{\sum_{t \geq 1} \mathcal{D}_{KL}(q_{x_{t-1}|x_t, x_0}||p_\theta(x_{t-1}|x_t))}_{\mathcal{L}_{consistency}} - \underbrace{\log p_\theta(x_0|x_1)}_{\mathcal{L}_{reconstruction}} \Big]
\end{aligned}
\tag{7.21}
$$

Where the Eq. 7.21 denotes the *Variational Bound* for $\log p(x)$ with the non-negative KL terms. Therefore, maximizing the log-likelihood of the prediction is equivalent to simultaneously minimizing KL terms. The $\mathcal{L}_{consistency}$ term uses KL divergence to compare the predicted reverse process $p_\theta(x_{t-1}|x_t)$ against the simulated reverse process $q(x_{t-1}|x_t, x_0)$ with condition on x_0.

$\mathcal{L}_{\textbf{prior match}}$**:** The prior match term can be negligible due to the fact that the forward process q has not learnable parameter, since all Gaussian parameters are scheduled (as in Eq. 7.19).

$\mathcal{L}_{\textbf{reconstruction}}$**:** This loss term constrains the Gaussian bridges between the probability of the last time points $(x_t|_{t=1})$ in the reverse process and the ones of training data x_0. Therefore, its probability is replaced by $\mathcal{N}(x; \mu_\theta(x_1, 1), \sigma_1^2)$, and the log-likelihood can be written as $\frac{-1}{2\sigma^2}\|x_0 - \mu_\theta(x_1, 1)\|^2 + C$. Eventually, constant and the variance terms can be ignored.

Therefore, the overall objective function for the diffusion training process can be simplified as computing the KL divergence between the posterior of the forward process and the learned reverse process. Knowing the Gaussian nature of both Eq. 7.19 and Eq. 7.14, the KL divergence can be expressed as in Eq. 7.23.

$$\mathcal{D}_{KL}(q||p) = \mathcal{D}_{KL}(\mathcal{N}(\mu_q, \Sigma_q)||\mathcal{N}(\mu_p, \Sigma_p)) \tag{7.22}$$

$$= \frac{1}{2}\Big[(\mu_q - \mu_p)^T \Sigma_q^{-1}(\mu_q - \mu_p)\Big]$$

$$= \frac{1}{2\tilde{\beta}_t}\|\mu_q - \mu_p\|_2^2 \tag{7.23}$$

$$= \frac{1}{2\tilde{\beta}_t}\frac{\hat{\alpha}_{t-1}\cdot\beta_t^2}{(1-\bar{\alpha}_t)^2}\mathbb{E}_{q(x_t|x_0)}\Big[\|\hat{x}_\theta(x_t,t) - x_0\|_2^2\Big]$$

Substituting in Eq. 7.18

$$= \frac{1}{2\tilde{\beta}_t}\frac{(1-\alpha_t)^2}{\alpha_t(1-\bar{\alpha}_t)}\mathbb{E}_{q(x_t|x_0)}\Big[\|\hat{\epsilon}_\theta(x_t,t) - \epsilon_t\|_2^2\Big]$$

ignoring the weight as was shown empirically by [12], the KL divergence of the above can be simplified as:

$$= \mathbb{E}_{q(x_t|x_0)}\Big[\|\hat{\epsilon}_\theta(x_t,t) - \epsilon_t\|_2^2\Big] \tag{7.24}$$

With the above derivation, the loss function can be eventually optimized using gradient descent. The model θ can be therefore updated via:

$$\theta_{i+1} \leftarrow \theta_i - \eta\cdot\nabla_\theta\|\hat{\epsilon}_\theta(x_t,t) - \epsilon\|_2^2 \tag{7.25}$$

with η being the step size of the gradient descent.

7.5 Conclusion

With the detailed derivation, it is less ambiguous to understand the design choices of the DDPM model. The three pillars of the model include: time-reversal Langevin dynamics for the forward and reverse processes, maximizing likelihood using a tighter bound, and the noise prediction loss function in the end. The field of diffusion models are still very active with rapidly evolving new variations of models, but the fundamentals remain the same. This chapter is written with intention to provide a comprehensive understanding of the DDPM model as well as the strong motivation to apply such models for MRI enhancement at a lager scale.

Chapter 8

Super resolution for MRI

8.1 Introduction

In this chapter, we study the learning-based approach for super resolution task of 3D MRI. Specifically, we investigated different types of generative models for restoring 3D high-resolution anatomical detail of MRI using adversarial training.

High-resolution MR images are crucial for obtaining a detailed view of anatomical structures and are essential for downstream MRI analysis. However, the acquisition of high-resolution (HR) MRI is labor-intensive and susceptible to motion artifacts during the scan. On the other hand, reducing scan time often results in lower spatial resolution and impairs fine details in the image structure. Therefore, many researches in deep learning have emerged towards the single image super resolution (SISR) task, which aims to restore low-resolution MR images to high-resolution images of the same subject.

In this chapter, we propose a new method for 3D SISR using a GAN framework. Methodologically, our approach incorporates instance noise to improve the training stability of the GAN, a relativistic GAN loss function, and a concurrently updating feature extractor to encourage perceptual generation during the training process. Eventually, our method is demonstrated to produce highly accurate and sharp results using very few training samples. Trained on patches, the model only require less than 30 whole-brain volumes, which is a significant reduction compared to the thousands typically required in previous studies. Moreover, our method excels in out-of-sample super-resolution task, demonstrating its robustness.

8.2 Methods

The pipeline of our method is shown in Fig. 8.1(a). The model is trained on paired low-resolution (LR) and high-resolution (HR) images, where the LR images are linearly down-sampled from HR images. The LR images are fed into the generator to generate super-resolution (SR) images, which are then evaluated by the discriminator. In addition to common GAN modules, the feature extractor is used to extract features from both SR and HR images, which are then compared to ensure consistency in the feature space. The discriminator is updated by the gradient from

adversarial loss, the generator is updated by the gradient from a combined loss of adversarial loss, pixel loss, and perceptual loss.

Model Design. Our model is based on the GAN architecture, with three learnable modules: a generator for generating realistic SR images, feature extractor to enhance feature space consistency, and a discriminator for the adversarial training. **Generator.** The generator is composed of 3 residual-in-residual blocks (RRDBs) [78] and a voxel-shuffle layer in 3D [79] for up-sampling the spatial dimension of the feature maps. **Discriminator.** The discriminator is composed of the output from a critics network subtracted by the mean of the counterpart sampled. In this setup, a critics is a common discriminator without the last non-linear activation function before output. Here, the two terms are used interchangeably to refer the same model. The discriminator contains four ResNet blocks comprised of stride convolution layers followed by instance normalization and Leaky ReLU activation functions. The numbers of the channels of the input and output are the same for neighboring blocks, except for the initial block which has to match the input image channel. A convolution layer with output channel equals one is added to restrict the final output. **Feature extractor.** The feature extractor uses lower layers of a 3D ResNet-10 before linear layers, without pretrained weights. All three networks are trained concurrently, where Gaussian noise is added to both real and generated inputs of the discriminator. The scheduled noise annealing ensures training stability and convergence to the optimal point.

Choice of layers. As building blocks of the generator, we implemented three RRDBs [78, 80]. Deep architectures can be prone to instability during training, where gradient attenuates as the model getting deeper. Especially in a GAN model, gradients are prone to vanish or explode. Thus, RRDBs uses multi-level skip-connection to improve computational efficiency and performance while retaining training stability [17, 14]. Feature matching has been shown to improve the perceptual quality of images in previous studies [81, 11]. In contrast to the standard perceptual loss which uses pretrained VGG19 network for feature extraction [11, 17, 14], we found that ResNet10, a shallower architecture, performed better in our experiments when trained from scratch.

Balancing the GAN. We add linearly annealed Gaussian noise [82] to both the SR and HR inputs of the discriminator during training, to avoid early phase divergence by forcing these two distribution overlapped. The Gaussian noise is from a standard normal distribution with its variance linearly decreasing per iteration from 1 to 0. This has been shown to be an effective and computationally cheap way of balancing the GAN dynamics and ensuring convergence [83, 82].

Optimal point. The RaGAN loss [83] objective function is designed to converge to an optimal point, where the discriminator $D_{Ra}(\cdot,\cdot)$ is optimized to predict a higher probability for the real sample x to be more realistic than fake sample y, and the generator is encouraged to generate fake samples that are more realistic than randomly drawn real sample. Here x is a mini-batch of samples from the real sample distribution $\mathbb{P}$, and y is sampled from fake sample distribution $\mathbb{Q}$, with $\tau(\cdot)$ being the *sigmoid* activation function.

The optimal point is reached when both the generator and discriminator can not

further improve themselves, meaning the probability estimated by the discriminator D_{Ra} for real-fake pair and fake-real pair equals (as in Eq. 8.4).

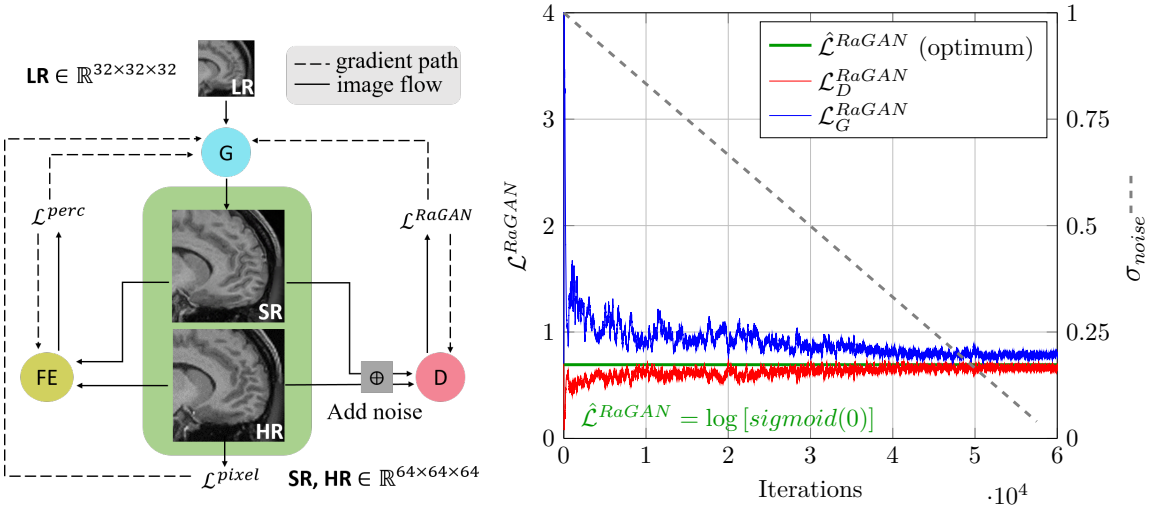
$$\begin{aligned} D_{Ra}(x, y) &= D_{Ra}(y, x) && (8.1)\\ &= \tau[C(x) - \mathbb{E}_{y\sim\mathbb{Q}}(y)] && (8.2)\\ &= \tau[C(y) - \mathbb{E}_{x\sim\mathbb{P}}(x)] && (8.3)\\ &= \tau(0) && (8.4) \end{aligned}$$

With optimally converged discriminator and generator, the adversarial loss ($\mathcal{L}^{RaGAN}$) also converges when the relativistic average discriminator yields a numeric value of $D_{Ra}(\cdot, \cdot) = \tau(0) = 0.5$ at the equilibrium point.

$$\mathcal{L}_G^{RaGAN} = -\mathbb{E}_{x\in\mathbb{P}}[\log(1 - (\tau(0))) - \mathbb{E}_{y\in\mathbb{Q}}[\log(\tau(0))] \quad (8.5)$$

$$\mathcal{L}_D^{RaGAN} = -\mathbb{E}_{x\in\mathbb{P}}[\log((\tau(0))) - \mathbb{E}_{y\in\mathbb{Q}}[\log(1 - \tau(0))] \quad (8.6)$$

Evaluating the converged value in each loss function, the binary cross entropy, for being either real for discriminator or fake for generator, can be rewritten as $-\log\tau(0) \approx 0.693$. As is shown in Fig. 8.1(b), the adversarial losses of both the generator and discriminator converge to the optimal value as in Eq. (8.6), when trained under annealed noise (gray dashed line).



(a) (b)

Figure 8.1: Training pipeline of the model. (a) Schematic diagram of our network during training. *D*, *G*, *FE* denote *discriminator*, *generator* and *feature extractor* respectively. Low resolution volumes (LR) are fed to the generator to produce super-resolution images(SR), the high resolution images (HR) and SR are distinguished by the discriminator. (b) Convergence of the adversarial losses during training, see its result in Fig. 8.3(e). Both $\mathcal{L}_D^{RaGAN}$ and $\mathcal{L}_G^{RaGAN}$ converge to their theoretical optimum (noted as $\hat{\mathcal{L}}^{RaGAN}$ in green line), under the annealing noise variance (σ_{noise}).

8.2.1 Training

Adversarial training is the core mechanism in the GAN training. The discriminator is used for a overall evaluation of the generator, with curated design choice for task-specific purpose. As input for the GAN model, the whole-brain volumes are randomly patched into mini-batch to compute the losses.

Objective Functions. In order to ensure both high perceptual quality of generated images and balanced training dynamic, we include a pixel loss ($\mathcal{L}^{pixel}$) and a perceptual loss ($\mathcal{L}^{perc}$), namely l_1 losses in image and feature domain respectively, and a GAN loss as the objective function for the generator network. For generator, the overall loss is defined as:

$$\mathcal{L}_G = \mathcal{L}^{perc} + \alpha\mathcal{L}^{pixel} + \beta\mathcal{L}_G^{RaGAN} \tag{8.7}$$

where hyper-parameters $\alpha = 0.01$ and $\beta = 0.005$ are as default in similar work [14]. The parameters of the generator are updated by Adam [84] optimizer with respect to $\mathcal{L}_G$. The discriminator and the feature extracting network are updated based on their respective losses, $\mathcal{L}_D^{RaGAN}$ and $\mathcal{L}^{perc}$, as defined in the following.

Our perceptual loss is defined as a loss in feature space, see Eq. 8.8, which simultaneously updates the feature extractor and generator. Here $I_{i,j,k}$ and $\hat{I}_{i,j,k}$ represent HR and SR intensity values at the i, j, k-th voxel in all three image dimensions (W, H, D), $\mathcal{F}(\cdot)$ represent the feature extractor network.

$$\mathcal{L}^{perc} = \frac{1}{W \times H \times D} \sum_{i,j,k=1}^{W,H,D} |\mathcal{F}(I_{i,j,k}) - \mathcal{F}(\hat{I}_{i,j,k})| \tag{8.8}$$

8.2.2 Inference

The inference process includes three steps: patching, generating, and assembling. Patching takes LR volume as input and uses the same size of $32 \times 32 \times 32$ as in the training as default, only the overlapped steps are changed to zero. In the generating step, we pass the patched volumes as mini-batch to the generator with frozen parameters inside the model. Eventually, all the patches upsampled by the neural network are assembled accordingly into an SR whole-brain volume.

Localized convolution operations. As a basic concept of convolution neural networks, "convolution operation" is simplified as a sliding matrix multiplication between input images and convolution kernel with the fixed perception field. Equally, the sliding-window computation allows the trained CNN to operate on input of any size larger than the perceptual field of convolutional kernel, as in Fig 8.2. Practically, this is helpful for inference on entire volume instead of patches when GPU memory is limited and using CPU memory for inference is feasible, which is commonly larger than GPU memory.

A simplified notation of the inference-time convolution operation is as follows:

$$\hat{X} = W_{m,n} * X \tag{8.9}$$

$$= \sum_{i,j}^{H,W} W_{m,n} * x_{i,j} \tag{8.10}$$

$$\tag{8.11}$$

where generated entire volume $\hat{X} \in \mathbb{R}^{B,C,H,W}$ can be obtained by either directly multiplying model weights $W_{m,n}$ on the entire volume $X \in \mathbb{R}^{B,C,H,W}$ (Eq. (8.10)) or, equivalently, the summation over multiplication between weights and patched volumes $x_{m,n} \in \mathbb{R}^{B,C,\frac{H-s}{k},\frac{W-s}{p}}$ (Eq. (8.11)); step size s and numbers of patches k, p are selected such that the size of the entire brain volume H, W can be dividable by the patch size m, n.

8.2.3 Implementation details

During the training of the network, we simultaneously trained a generator, a critic, and a feature extractor network. To fit in the GPU memories, we crop the whole brain volumes into overlapping HR patches of size $64 \times 64 \times 64$. The LR patches were obtained by linearly down-sampling HR patches to a matrix size of $32 \times 32 \times 32$. The paired LR and HR patches were fed into the model for training.

All three networks are initialised by Kaiming initialization [85] and are optimised by Adam optimizers [84], using the default coefficients of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate of $1e^{-4}$. The variance of the instance noise is scheduled to decrease linearly per epoch for 20 epochs, from $\sigma = 1$ to 0. The training is implemented using the PyTorch framework [86] and run for 60000 iterations, on an NVIDIA's Ampere 100 GPU.

Evaluation Metrics. To quantify SR quality, we use Peak-Signal-Noise-Ratio (PSNR), Structure-Similarity-Index-Measurement (SSIM), and Learned-Perceptual-Image-Patch-Similarity (LPIPS) [87]. However, compared with the other metrics, LPIPS better reflect image fidelity [88, 89]. Here we use the 2D slice-wise LPIPS implementation provided by [87], which has been pretrained on large image datasets and are known to extract meaningful features that closely resemble human visual quality [88, 87].

8.3 Experiments

Datasets. We used publicly available datasets from different scanners, imaging modalities or parts of body. Three of these datasets are brain MRI acquired at 3 T, which are part of the Human Connectome Project [90], the other one is a knee dataset with proton density (PD) contrast at 3 T from [91]. We name the datasets by the most distinguishable attribute of the each dataset, as described in the table 8.1 below:

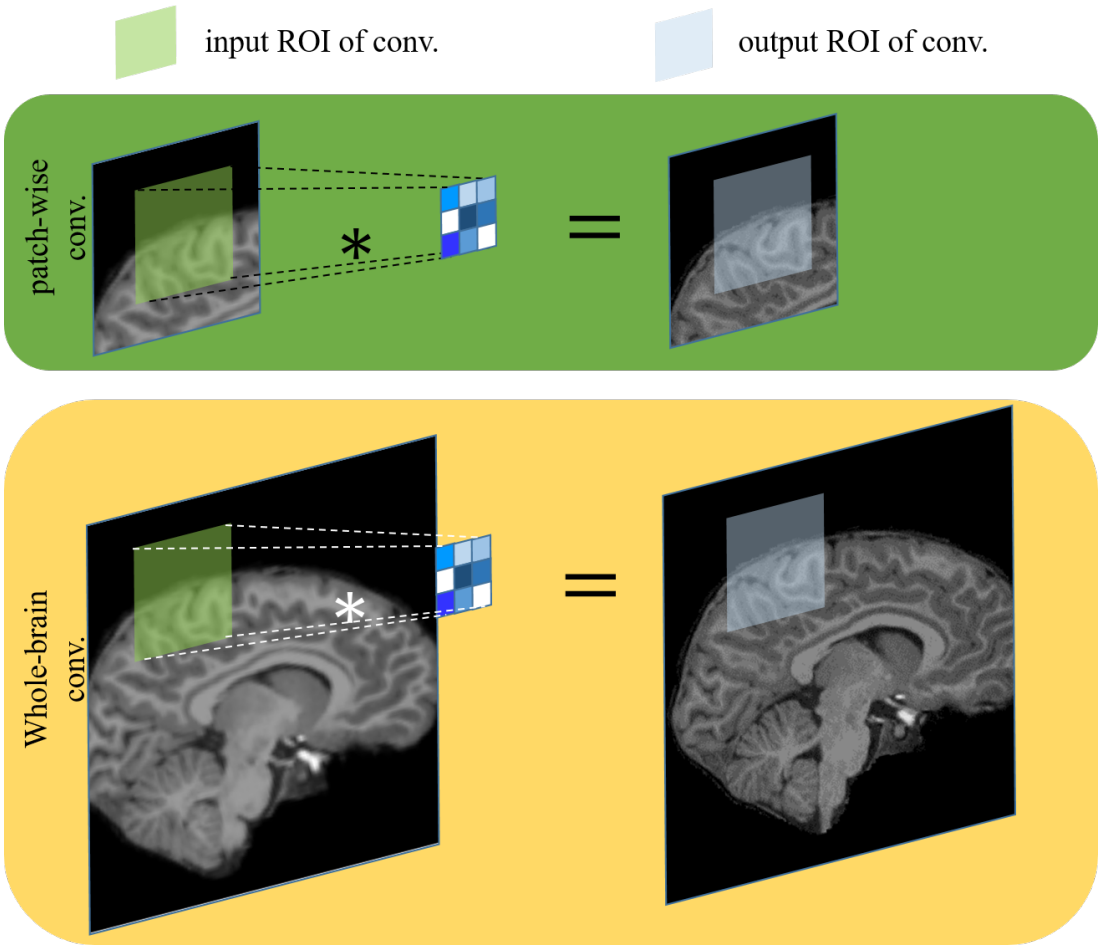


Figure 8.2: Equivalence of model inference on patch and entire volume. Inference time convolution operation outputs identical intensity values on given perceptual field for both patch and whole image input. The output image is acquired by sliding the kernel over the input image, namely the convolution operation, with the output size determined by the kernel size and the stride at each step. Thus, the eventual output is not effected by the input size, but only by the coefficients of the kernel and its size and stride.

Table 8.1: All datasets used in the experiments

Dataset name	Dataset source	Resolution	Contrast	Number of subjects	Vendor
"Insample"	*Lifespan Pilot Project*(HCP [90])	0.8mm	T1w	27	Siemens
"Contrast"	*Lifespan Pilot Project*(HCP [90])	0.8mm	T2w	27	Siemens
"Resolution"	*Young Adult study*(HCP)	0.7mm	T1w	1113	Siemens
"Knee"*	*Fully sampled knees* [91]	0.5/0.6mm	PD	20	GE clinical

Table 8.2: Quantitative comparison of robustness among the *state-of-the-art* models on ***Dataset "Insample"*** and ***Dataset "Resolution"*** . Note the LPIPS score matches perceptual quality better than PSNR and SSIM, as shown in Fig. 8.3 and Fig. 8.5

Dataset names	Models	PSNR ↑	SSIM ↑	LPIPS↓
Dataset "Insample"	Tri-linear	33.038	0.876	0.084
	ESRGAN [92]	37.022	0.933	0.044
	DCSRN [93]	**37.635**	**0.954**	0.052
	ArSSR [94]	28.038	0.280	0.291
	ours	36.922	0.943	**0.037**
Dataset "Resolution"	ESRGAN	37.181	0.957	0.039
	DCSRN	**37.564**	0.962	0.051
	ArSSR	36.550	**0.970**	0.055
	ours	36.922	0.953	**0.038**

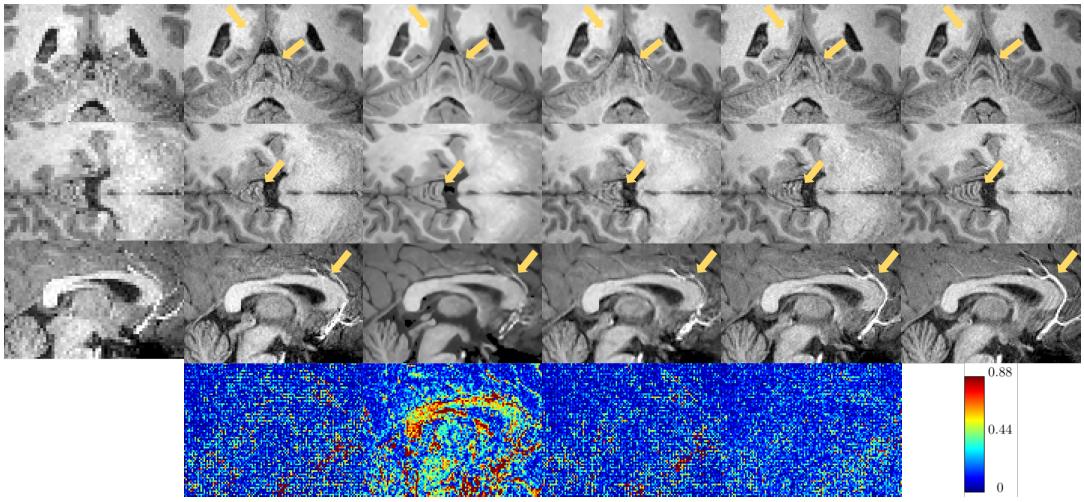
8.3.1 Results

Baseline. For comparison, we retrained ESRGAN, ArSSR, DCSRN for fair comparison MRI super resolution in 3D. These models are based on GAN and implicit neural representations, which were shown to be effective on super resolution task. However, our method combining the favorable optimization property and effective latent representation, has shown *state-of-the-art* results.

In-sample super resolution. As is shown in Fig , our model recovered most details with smallest residual against the ground truth. Notably, the blood vessel, which is not semantically the major structure in the MRI, is distinguishably restored in our result. The quantitative results in table 8.2 also highlight the effectiveness of our method by statistical metric, PSNR and SSIM, and perceptual metric, LPIPS.

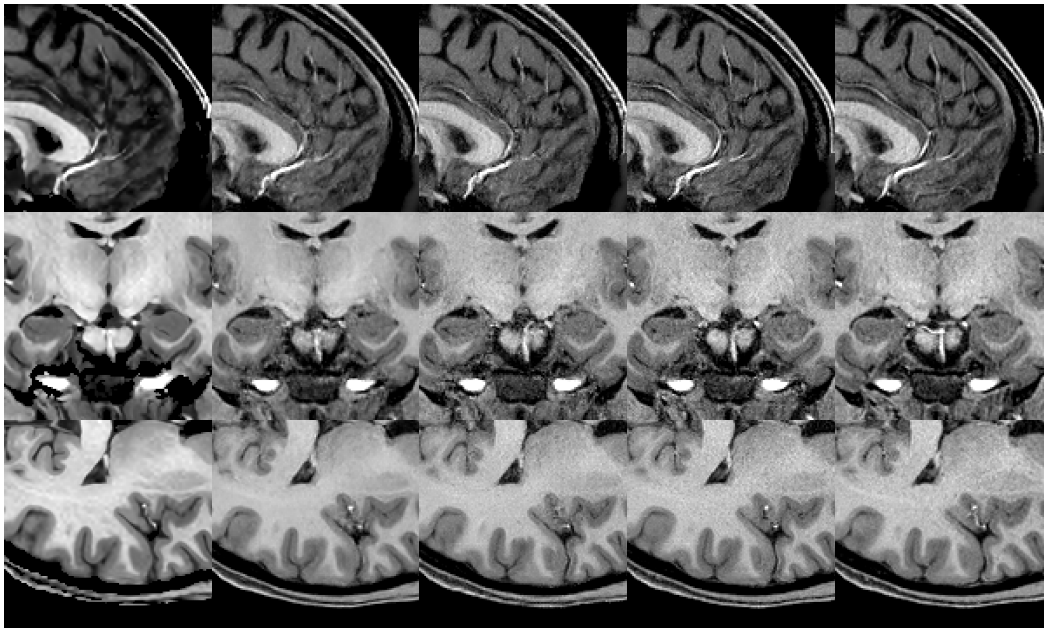
In addition, we noticed our model trained on dataset "Insample" also successfully super resolved MRI images from "Resolution", where the former has resolution of $0.8mm$ and the later has $0.7mm$, as in Fig. 8.3.

Tested on unseen datasets. Our model is further evaluated on an unseen dataset "Resolution", with similar content by slightly different resolution. As shown in Fig. 8.4, most of the models generate SR image successfully, but our model shows the closest approximation for details, such as blood vessel and minimal modification in the brain structure. This validation demonstrates the generalizability of our model to unseen datasets, without major decrease in performance.



(a) LR (b) ESRGAN (c) ArSSR (d) DCSRN (e) Ours (f) GT

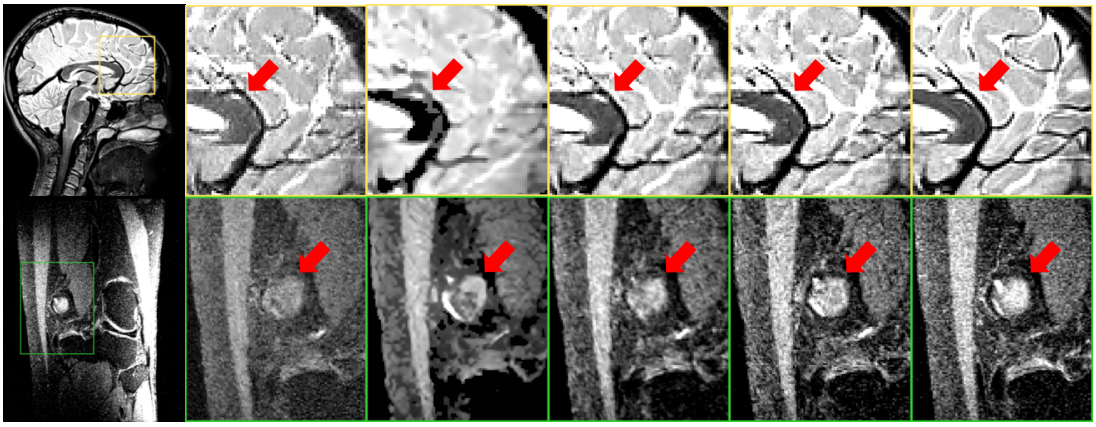
Figure 8.3: Qualitative comparison on an in-sample subject using different models with a $\times 2$ resolution upscale. ***Top*** three rows show MRI images from different models, the ***last*** row shows corresponding residual maps between SR images and the ground truth (GT). **(a)** shows the zoomed-in views of low-resolution (LR) images of brain MRI GT; The zoomed-in views of various SR results are shown in **(b)-(e)** and the GT in **(f)**. Our results show the best visual quality, the most detailed recovery of the GT, and the smallest residuals in brain structure. Distinguishable differences are marked by *yellow* arrows.



(a) ArSSR (b) DCSRN (c) ESRGAN (d) Ours (e) GT

Figure 8.4: Qualitative comparison of model performance using dataset "Resolution" showing zoomed-in SR results with different perceptual quality. (a)-(e): ArSSR, DCSRN, ESRGAN, **Ours**, and GT. *Note*: Our result shows the best visual quality and the lowest LPIPS score, despite of having a low PSNR and SSIM score.

Out-of-distribution super resolution. To test the generalizability of our method, we evaluate the model trained on dataset "Insample" on dataset "Contrast" and "Knee". The ground truth data from "Contrast" and "Knee" dataset shows significantly different from the training dataset, in terms of contrast, body parts. As is shown in Fig. , our model best recovered the anatomical details of the images, by restoring the blood vessels and details bone structures.



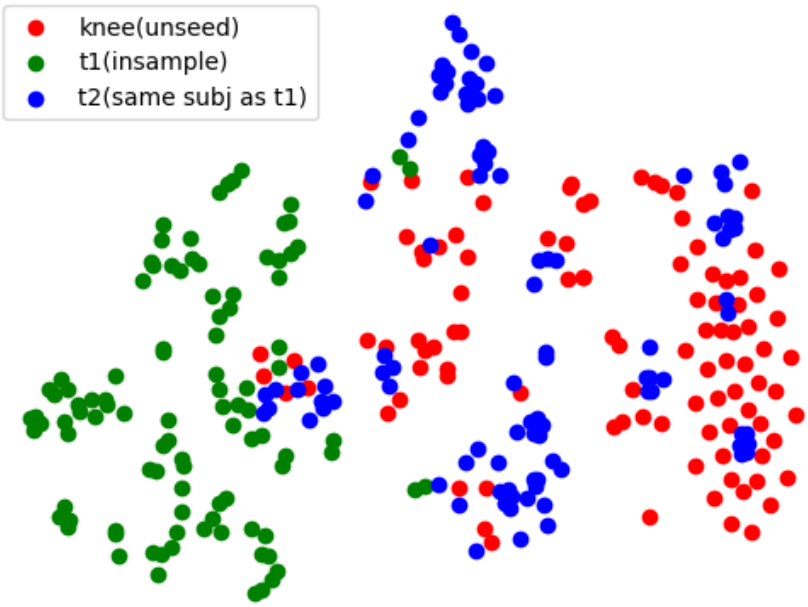
(a) (b) (c) (d) (e) (f)

Figure 8.5: Qualitative comparison on out-of-distribution datasets "Contrast" (*top row*) and "Knee" (*bottom row*) for different models. (a) shows the overall view of GT image and (b)-(f) are zoomed-in SR from corresponding methods; distinguishable differences are marked in *red* arrows

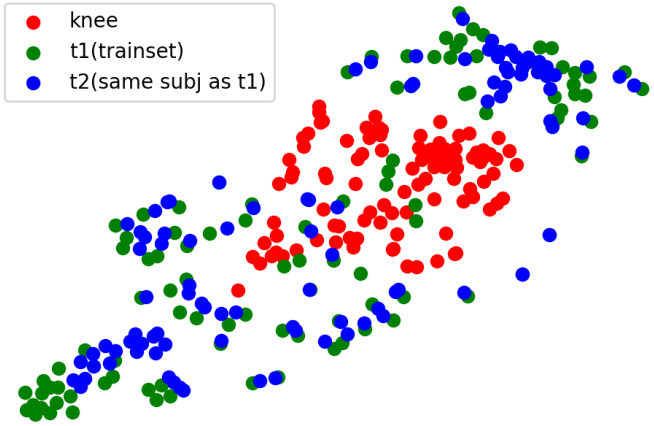
8.3.2 Generalized feature extraction

To further investigate the effectiveness of the feature extractor module (FE), we compare the data representation and the latent representation of the FE in low dimensional space using T-SNE. The visualization is shown in Fig. 8.6. As a comparison, tighter and more structured clustering of the data points in the latent space is observed in the FE output, as shown in Fig. 8.6(b). Moreover, as the T1w and T2w images are from the same subject, the overlapped clusters between T1w and T2w suggest their highly overlapped structural similarity. However, low-dimensional manifold of data representation distribute more sparsely without semantically meaning overlapping. This indicates that the FE module can be generalizable to different modality and body parts, helping the GAN to avoid falling into mode collapse (only memorizing image patterns exist in training data).

It is worth to notice that the T1w and T2w images are from the same subject, thus, their latent representations of FE show heavily overlapped clustering. However, the FE output of the "Knee" dataset shows a more excluded manifold, which is consistent with the visual semantics of the SR images in Fig. 8.5.



(a) input data representation



(b) latent representation of FE

Figure 8.6: Low-dimensional feature representation by T-SNE of data representation and latent representation of the feature extractor. Note the semantically assembled clusters for T1w and T2w(same subject as T1w) are only shown after the feature extractor, and are distant from the knee MRI, indicating the effectiveness of pattern extraction.

8.4 Conclusion

In this chapter, an effective method is proposed for stabilizing GAN training, which outperforms *state-of-the-art* SR models on 3D brain MRI data. Our approach involves adding Gaussian noise to discriminator inputs, forming a three-player GAN game, and applying the relativistic GAN loss as the objective function. Our experimental results demonstrate that this method can ensure effective training convergence and achieve better generalizability. Several key findings are presented in this paper. First, the proposed model generates SR results with superior perceptual quality compared to existing methods, with very limited amount of training data. Secondly, the model achieves convergence using a simple trick of adding noise. Finally, the model exhibits strong generalizability and is less prone to overfitting, as evidenced by its successful performance on previously unseen datasets.

While the LPIPS score better reflects the perceptual quality of images, it is primarily designed for natural 2D images and may overlook information in the cross-plane of 3D MRI images. Therefore, a specialized metric for evaluating the perceptual quality of 3D MRI images is of high importance in the future.

Chapter 9

Super resolution with denoising for MRI

9.1 Introduction

High-resolution MR images (HR) play crucial roles in providing detailed anatomical information. However, the acquisition process always requires a curated protocol to ensure the subjects being scanned present minimum motion, which is usually not guaranteed in real-world cases. Various types of artifacts show up in real-world scans, where thermal noise and motion artifacts show up, degrade the image quality, and lead to lower spatial resolution and loss of anatomical details. Furthermore, when it comes to applications in low-field MRI, this can be an essential problem where images are acquired in lower resolution (LR) and more noisy content.

To this regard, we invested and developed a deep learning model for efficiently tackling both super-resolution (SR) and denoising tasks in one model, without additional training for each individual tasks. The result suggested our model effectively recovered the high-resolution reference images while applying noise removal regardless of input noise levels and types.

Conventionally, these two tasks are treated separately. SR method aims to recover HR MR images from LR images of the same subject, while denoising tasks focus on removing common noise sources such as Gaussian noise or motion artifacts to obtain cleaner image content. Normally, these tasks require separate training and paired datasets in most deep-learning methods. Here we propose a frequency informed GAN model, DISGAN, which simultaneously performs SR and denoising tasks in one model. The model is trained only on paired HR and LR images as most of the SR tasks, and hierarchical wavelet transformation is embedded into the discriminator to guide the generator to output images with high-frequency fidelity and ignoring redundant noise. The model is evaluated on real-world MRI data with different types of noise at different levels and compared with traditional denoising methods. To this end, we aim to provide an essentially SR model, leveraging the wavelet transformation for removing the noisy artifacts which are generalisable to real-world MRI data.

9.2 Methods

Model design. The architectural design of our neural net enables particular patterns of information to be learned for the task. We build our model based on GAN architecture, with a generator composed of Residual-in-residual blocks [78], a feature extractor, and an UNet [95] as a frequency-informed discriminator.

Based on the UNet [95] architecture, we append learnable convolution blocks after discrete wavelet transformations (DWT) in 3D for the discriminator. Subsequently, the feature representations, which are downsampled from "Resblock" and the ones downsampled from the low-frequency output from DWT blocks, are combined to be compressed; whereas, the upsampled feature map together with high-frequency coefficients from the DWT blocks is fused to reconstruct the final output (as is shown in Fig. 9.2). This design of architecture possesses the ability to omit trivial information, such as artifacts, but preserve prominent structures.

Unet vs. Resblocks.

Throughout the experiments, we have observed the "Resblocks" preserve information of the input image with less information compression, compared with Unet. In line with this, we implemented "Resblocks" for generator and Unet for the discriminator where the information-preserving feature of "Resblocks" is important for the generator's objective. Unet is deployed as a fundamental architecture for the discriminator where compressing frequency information hierarchically to match the semantics in GT is desired.

Discrete wavelet transform in 3D. We propose the discrete wavelet transform unit (DWT+conv) for extracting and parsing frequency-wise information in feature space. As a fundamental operation, *3D Haar wavelet transform* is used for the computation of sub-band coefficients, where *LLL* is the low-frequency output and the sum of the rest is the high-frequency output (*e.g. HLL, LHL, LLH, HHL, HLH, LHH*, and *HHH*). 1x1 convolution layers are added after each sub-band coefficient to learn sub-frequency features, as is shown in yellow and gray blocks in Fig. 9.1.

9.2.1 Training

The training of our DISGAN method is based on the 3-play GAN and is optimized by the same objective as Eq. (8.7), as stated in section 9.2. Except for that the frequency-informed discriminator, as shown in Fig. 9.2, is used in the the current method. As a comparison, we also implemented Swin-transformer for the basic architecture of the generator. The training procedure is carried out on patch-based volumetric MR images, using the same training dataset as was in Chapt. 8.

9.2.2 Inference

The inference for our model is the same process as in section 8.2.2. As a comparison method on noise removal task, we use the Rotation invariant non-local means filter(PRINLM), proposed in [96] for the denoising task. The PRINLM method is a

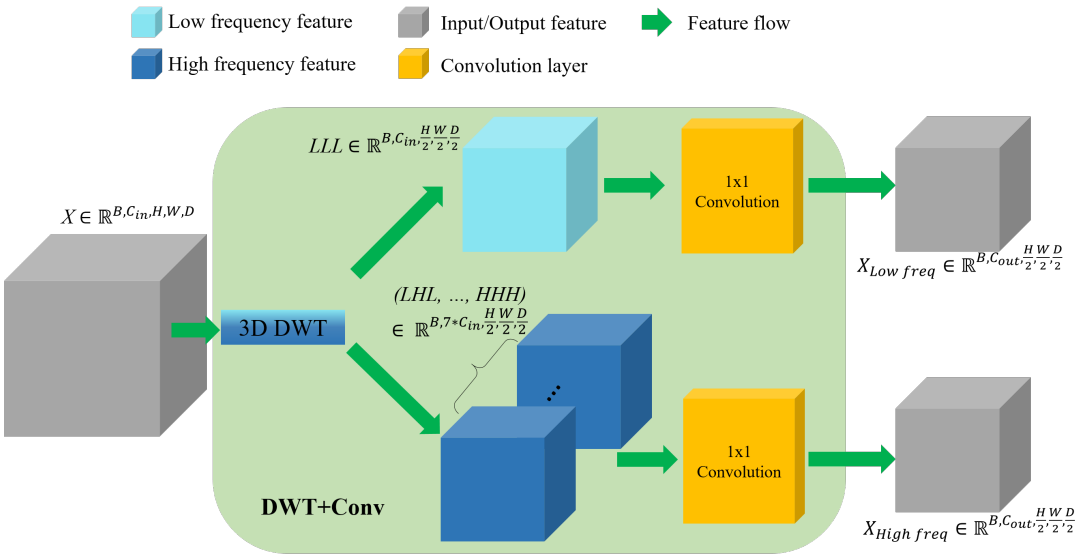


Figure 9.1: Schematic illustration of the DWT+conv unit. The input tensor is parsed into 8 sub-band coefficients by the 3D Haar wavelet transform, where the low-frequency and high-frequency coefficients are summed up and passed through two separate 1x1 convolution layers, resulting in low-frequency and high-frequency features.

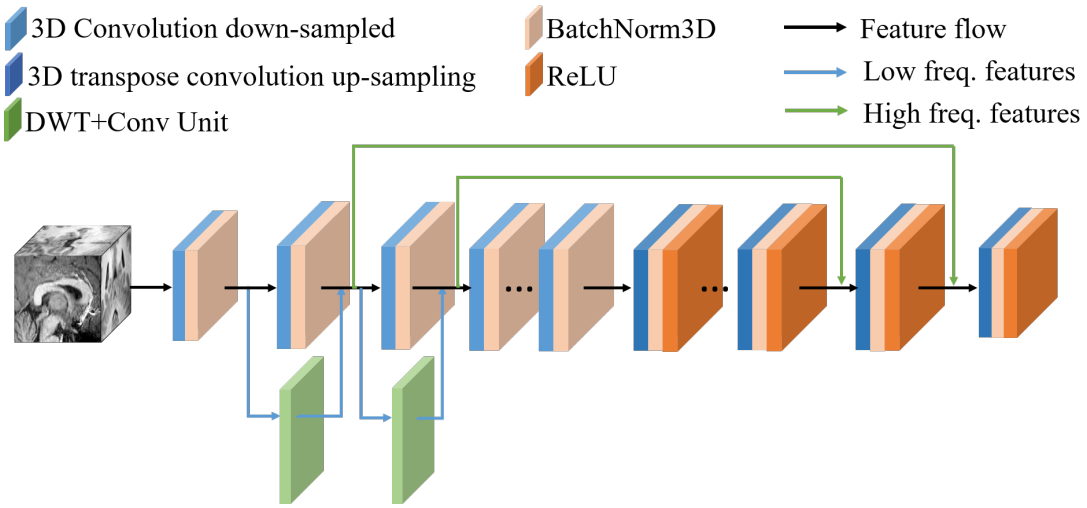


Figure 9.2: Schematic diagram of the frequency-informed Unet architecture in the discriminator. The network takes MRI patches in the high-resolution image space as input, and outputs voxel-level score as discrimination for *real* and *fake* samples.

traditional method for Rician noise estimation and removal, which is widely used in MRI image processing. The PRINLM method is based on the combination of discrete cosine transformation hard thresholding and non-local means filter, which removes the natural redundancy of the pattern within image. The PRINLM method is implemented in the MATLAB code [97].

9.2.3 Implementation

We trained a generator network, a critic network and a feature extractor network. The generator network consists of 3 residual-in-residual dense blocks [78], which are densely connected residual units embedded without a batch norm layer. The critic network is the same as a discriminator without the final nonlinear activation layer. The feature extractor uses convolutional layers of ResNet10 before linear layers. All three networks are initialized by Kaiming initialization [85] and optimized by Adam optimizer [84], using coefficients of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of $\gamma = 10^{-4}$. The variance of the instance noise decreases linearly in each iteration, from $\sigma = 1$ to 0. Training is done simultaneously on the PyTorch framework [86], the generator, and the discriminator network for 60000 iterations, on NVIDIA's Ampere 100 GPU (40GB).

9.3 Experiments

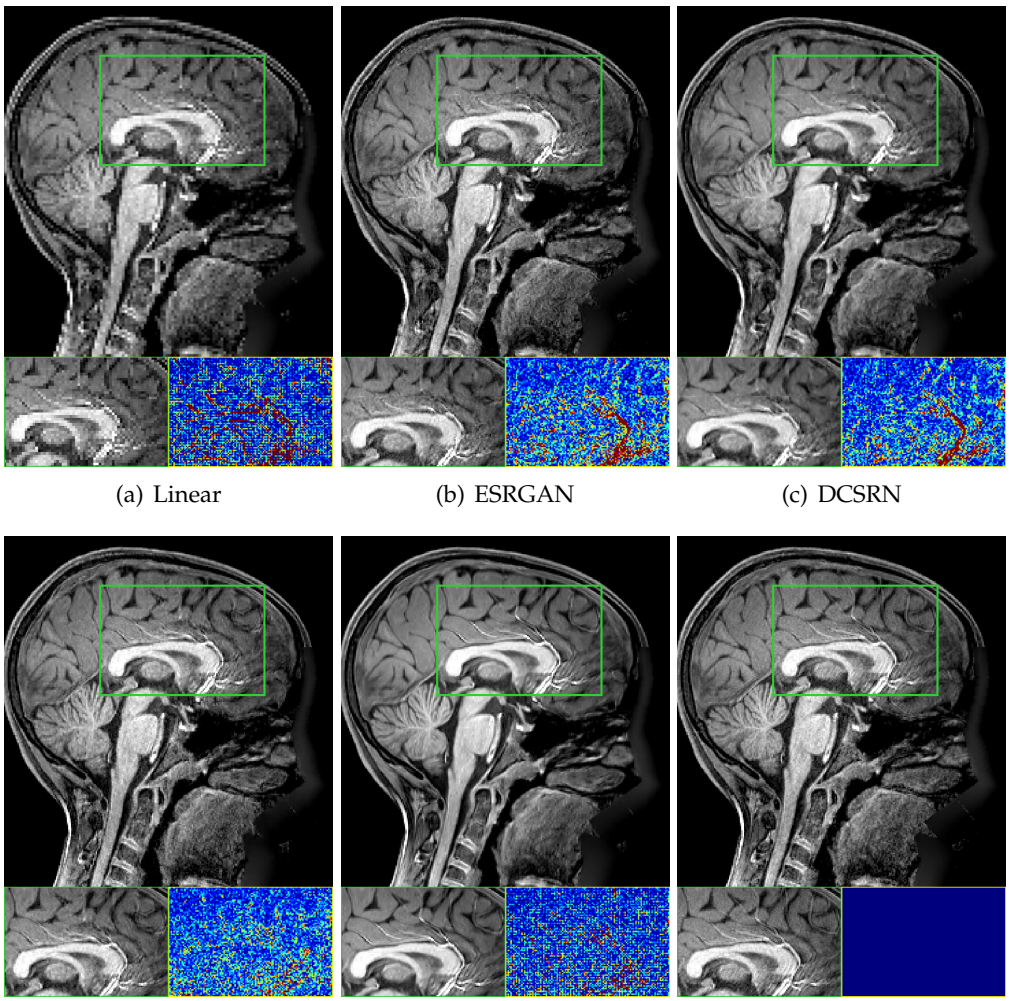
Datasets. We use publicly available datasets from different scanners, varied age groups and the health conditions of subjects. All training is performed on "Insample" and tested on "Epilepsy" and "Tumor", where the former contains brain images from epilepsy patients with salient motion artifacts and the latter from brain tumor patients with distinguishable tumors in the brains, from BraTS challenge. A detailed description of the essential information is given below (Table. 9.1):

Table 9.1: All datasets used in the experiments

Dataset name	Dataset source	Resolution	Contrast	Number of subjects	Vendor
"Insample"	*Lifespan Pilot Project*(HCP [90])	0.8mm	T1w	27	Siemens
"Epilepsy"	*OpenNeuro* [98]	0.8mm	T1w	9	Siemens
"Tumor"	*BraTS* [99]	1mm	T1w	9	Siemens

9.3.1 Results

Super resolution. In this experiment, we test the performance of our model against other 3D SR models on the *"Insample"* dataset. For comparison, we tested our approach against four other approaches, namely ESRGAN-3D [92], DCSRN, and Wang et al. [1]. ESRGAN was proposed in 2D, so we reconfigured it by replacing



(a) Linear (b) ESRGAN (c) DCSRN

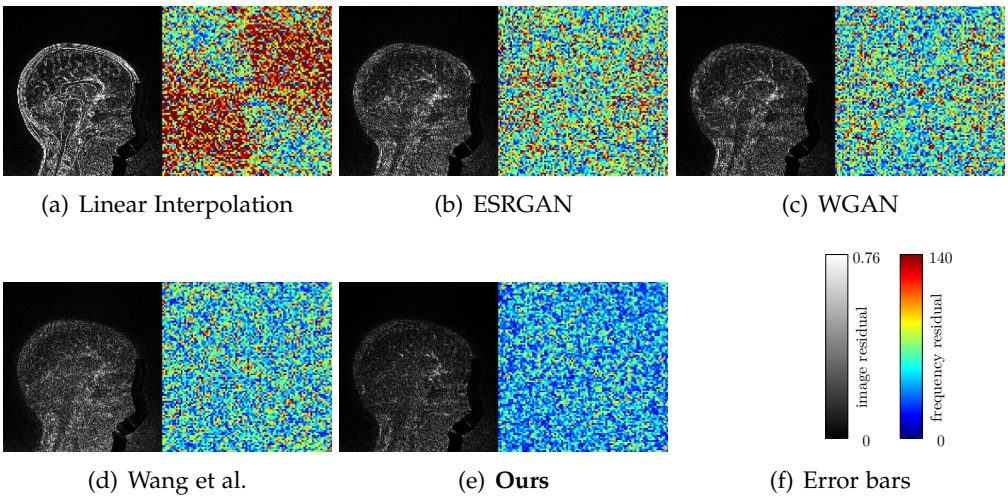
(d) Wang et al. (e) DISGAN (ours) (f) GT

Figure 9.3: Super-resolution results of an exemplary subject using various methods. The zoom-in view and the colormaps (red for high and dark blue for low residual, using jet colormap, same for all other colormaps) of the absolute residuals of the green boxes are shown below each column. Note that our method recovers detailed structures the most, such as the blood vessel (zoom in for a better view).

all 2D with 3D convolutions. Specifically, we pre-trained the first 54 layers of a 3D VGG-19 network on 3D MRI images from the same dataset. The DCSRN model is available in 3D, but we retrained the model on the "Insample" dataset.

The results are shown in Figure 9.3. Note that our model recovered the ground truth best, providing high-quality anatomical structures with less noise. This is particularly true for the detailed blood vessel structures, which are reflected by the smallest residual in the zoomed region; our model also performs dominantly by recovering the finest structures of the cerebellum. The quantitative results are shown in Table 9.1.

Furthermore, as shown in Figure 9.4, the DWT-informed discriminator of our model guides the generator to produce images with high fidelity in both image space and frequency space, which is obtained by Fourier transforming the magnitude image without explicitly constraining the frequency space data. We crop the center matrix of $50 \times 50 \times 50$ of the frequency space of the image for better contrast when visualizing the difference.



(a) Linear Interpolation (b) ESRGAN (c) WGAN

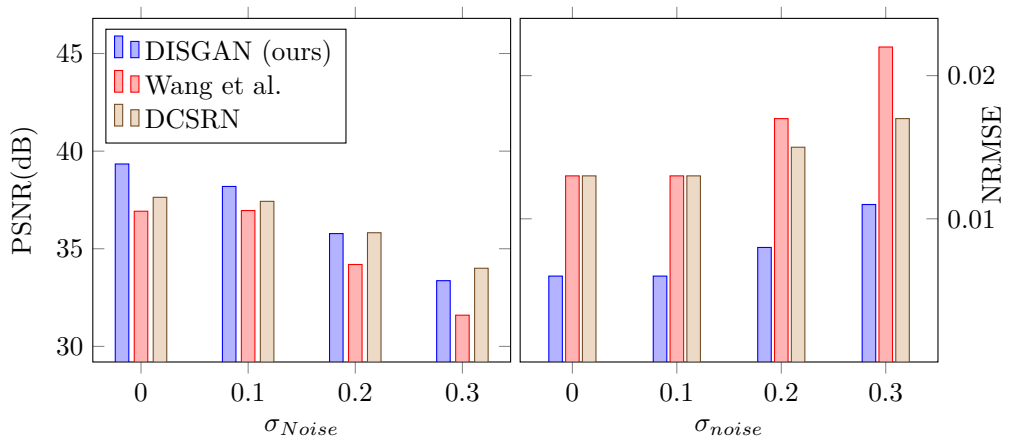
(d) Wang et al. (e) **Ours** (f) Error bars

Figure 9.4: Residual plots in image and spatial frequency domain of the same SR slices as in Figure 9.3 over different methods. Odd-numbered columns display the absolute residuals of SR methods against GT in image space, while even-numbered columns show the absolute residuals in the center-cropped frequency space for each model against the GTs. Note our model (DISGAN) exhibits the lowest residual in both image space and frequency space.

Cleaning simulated noise. To test the noise-removal utility of our model, we corrupted the GT images with simulated Gaussian noise with different levels of variance, to simulate the noisy data. All the models for comparison are dedicated to SR tasks and are evaluated on four different noise levels with a standard deviation of 0, 0.1, 0.2, and 0.3.

As is shown in Fig. 9.6, the GT image is corrupted with different levels of Gaussian noise and is downsampled, passed into different models for SR and noise-cleaning. It is obvious that the DISGAN preserves the most clear white matter/ gray matter boundaries throughout all results. Especially at the highest severity of noise, our DISGAN still preserves clean image content when $\sigma = 0.3$ (see green and yellow window of Fig. 9.6 in page 56). Moreover, the prominent score on traditional metrics, like PSNR and NRMSE, validates the statistical consistency of our model on removing simulated noise (see Fig. 9.5).

Cleaning simulated real-world noise. In this experiment, we test our model on restoring HR images from LR and removing noise concurrently on real-world data, where MRI samples with unseen noise are fed to the model. Specifically, brain MRI



(a) PSNR comparison (b) NRMSE comparison

Figure 9.5: Quantitative comparison of all models on PSNR and NRMSE metrics. Higher PSNR and lower NRMSE indicate better performance. Note that our model achieves the highest PSNR and lowest NRMSE, indicating the best performance in SR tasks at all noise levels.

data with tumor and epilepsy disease are used for this experiment. The image with the brain tumor is confounded by noisy tumor contour and random noise, from the dataset "Tumor" (as is in Fig. 9.7(a)); the image with epilepsy disease is contaminated by motion artifacts, presenting repetitive patterns (as is in Fig. 9.8(a)), from dataset "Epilepsy".

Comparisons are made among our model, popular and traditional denoising tool SANLM [43], and a dedicated Rician noise removing tool, PRINLM [96]. The qualitative result suggests the prominent quality of denoising result by our model in both tumor data (Fig 9.7(b), 9.7(c)) and epilepsy data (Fig 9.8(b), 9.8(c)). After being processed by our model, the image quality of both images is clearly improved, where the aforementioned thermal noise is significantly alleviated 9.7. The clean images show significantly clear gray and white matter boundary as well as distinguishable contour of the tumor. Likewise, the motion noise in the epilepsy image is effectively removed, as highlighted by the green arrow in Fig. 9.8. In the comparison with the PRINLM method [42], our model shows a better performance in not only removing random noise but also alleviating the ghosting artifacts in the epilepsy image.

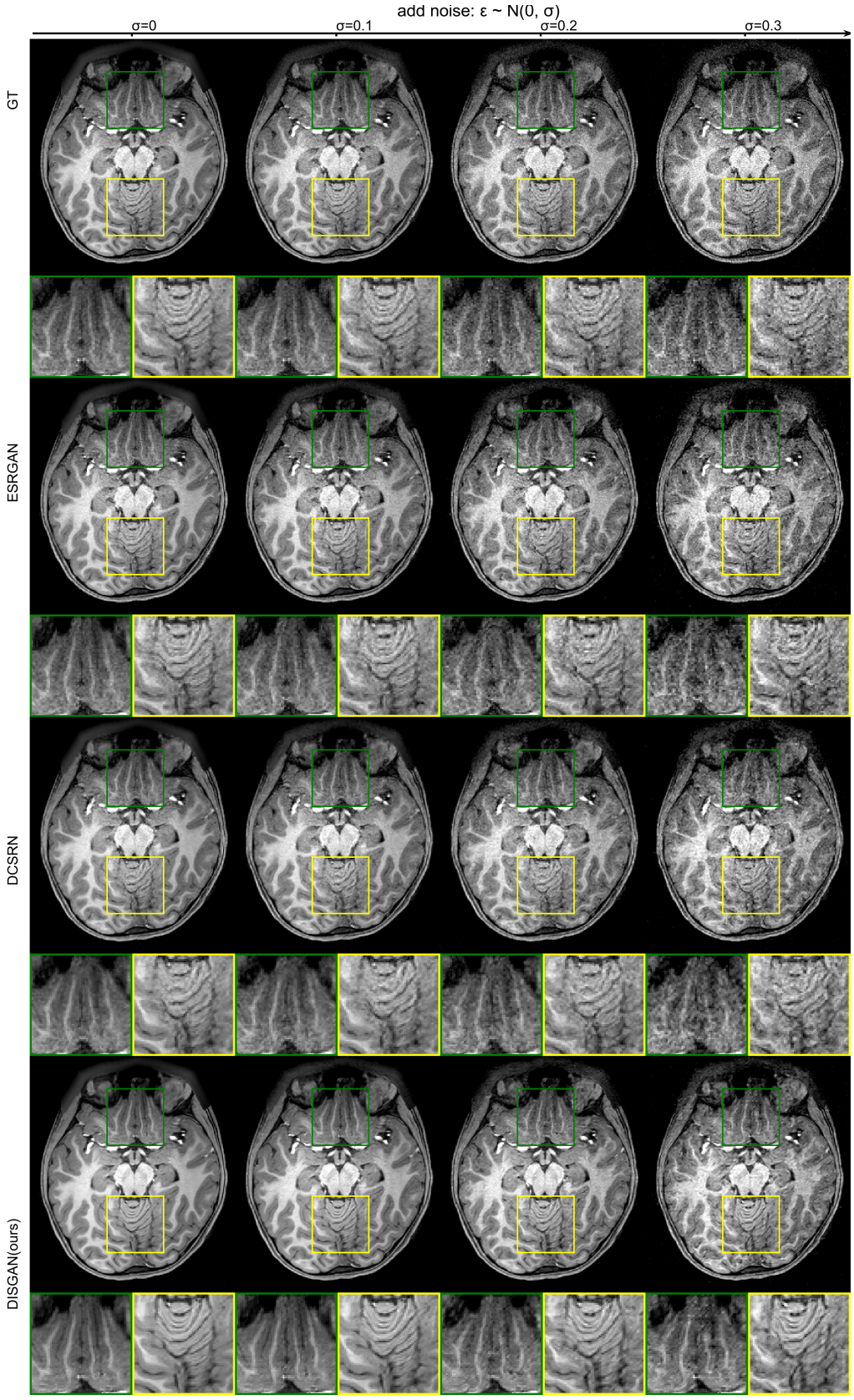
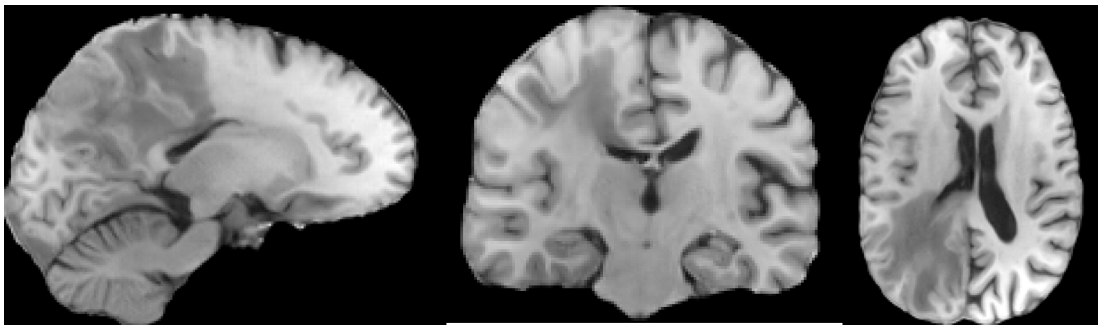
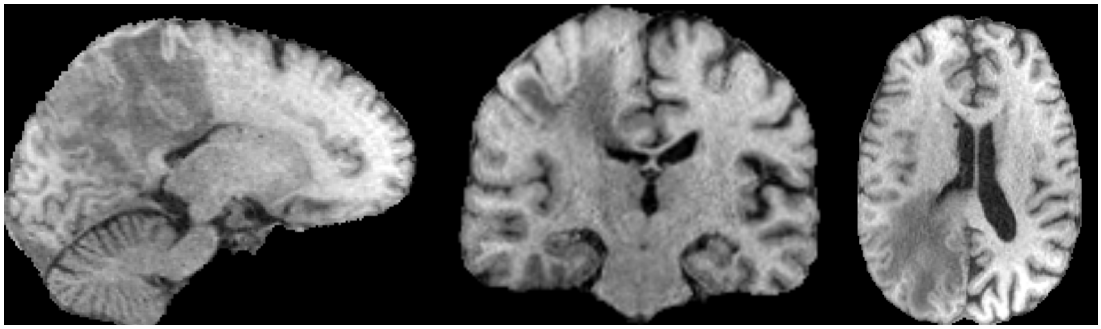


Figure 9.6: Simulated noise removal results from different SR models. Each column corresponds to generated images from different levels of noise corruption.

(a) Tumor (GT)



(b) Tumor denoised (**ours**)



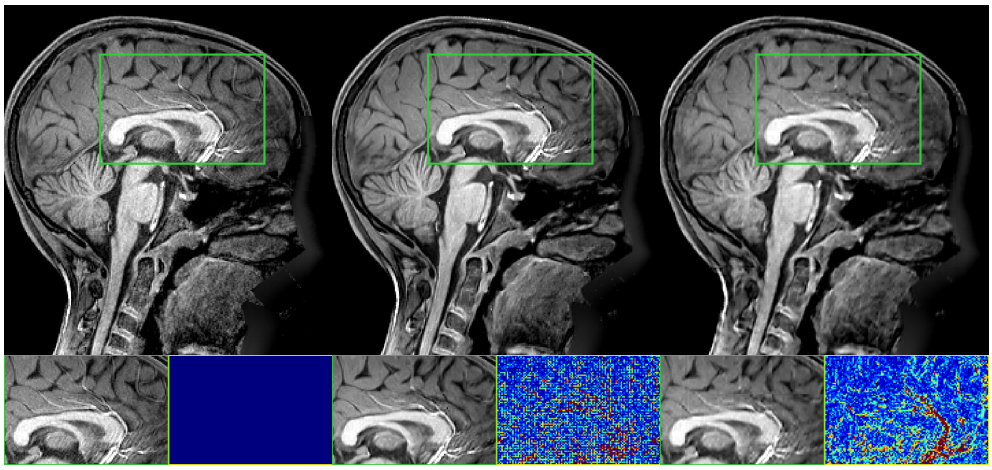
(c) Tumor denoised (PRINLM)

(d) Tumor denoised (SANLM)

Figure 9.7: Qualitative result for cleaning real-world noisy data with brain tumor. (a) shows reference images with tumor and random noise from the BraTS dataset; (b) shows images cleaned by **our** model; and (c)-(d) shows images cleaned by non-local filter based methods, PRINLM [42] and SANLM [43] respectively.

9.4 Ablation studies

Residual block vs. Transformers. We compare the performance of our model with the volumetric-residual-n-residual-blocks(VRRDB) and the Swin-transformer [100] as the building-block for the generator, with comparable numbers of parameters. Transformer architectures are dominant in almost all vision tasks, which motivates our evaluation to happen. However, the results show that the Swin-transformer is not as effective as the VRRDB in our model, as shown in Fig. 9.9 in page 58. The Swin-transformer fails to recover, as complete as VRRDB, the fine details of the blood vessels and the cerebellum, which are the under-represented structures in the brain.



(a) GT (b) DISGAN w/ VRRDB (c) DISGAN w/ SwinIR

Figure 9.9: Qualitative comparison between the output from DISGAN with VRRDB and SwinIR for generator building blocks. The zoom-in view and the colormaps distinguishes the fine details in the blood vessels by VRRDB.

9.5 Conclusion

In this paper, we propose DISGAN, a GAN architecture for SISR and denoising 3D MRI without the need for separate denoise training. Specifically, we propose an effective 3D DWT+conv block as a fundamental unit of our discriminator, which can indirectly guide the generator to output an image with high-frequency fidelity and minimal noise. Our experimental results prove that our DISGAN model achieves distinguishable SR results with the closest detail restoration while minimizing noise. However, it is still an open question to understand the design factor leading to the generalisability of GAN models. In the future, we also plan to investigate the maneuverability of DISGAN for dedicated denoising tasks and improve the denoising ability for such a unified model.

Figure 9.8: Qualitative comparison between different methods on real-world epilepsy data with strong motion artifacts. (a) shows GT images of epilepsy with motion noise (ghosting artefacts); (b) is noise-cleaned image from **our** model; and (c)-(d) are the images cleaned by PRINLM [42] and SAMLM [43] methods.

Chapter 10

Motion Correction for MRI

10.1 Introduction

In this chapter, we convey a study for retrospective motion correction using neural networks for anatomical MRI which is less studied compared to that in functional MRI. Motion artifacts are a common problem in MRI and can cause major degradation to image quality. It can be caused by patients' physical movement, physiological motion, or scanner instabilities. Motion artifacts can lead to image distortion, blurring, and ghosting, which affect the quality of the images and the accuracy of the diagnosis. Traditionally, motion correction techniques aim to reduce or eliminate these artifacts by aligning the images to a common reference image. However, this introduces labor intense work for implementing the hardwares and limits the generalisability to different vendors. In this chapter, we propose to use neural network-based the motion correction techniques for efficient image quality improvement for MRI. Neural networks are known for its effectiveness to learn the mapping between the motion corrupted images and the clean images, which naturally fits the aims of retrospective motion correction. Finally, we noticed this approach is applicable to in-distribution data, such as simulated motion images, but not well generalized to out-of-distribution data, such as real-world motion images which is under-represented by simulated data.

10.2 Methods

Motion simulation. We simulate the motion corrupted images by adding a combination of random translation and rotation to the ground truth images. Our motion corrupted images are generated by the TorchIO [101] package with a combination of slight and severe motion. The motion parameters are randomly sampled from uniform distribution, with the translation range of 10 pixels and the rotation range of 10 degrees for all 3 xyz-axis. The motion corrupted images and their paired reference images are then fed as the input to the motion correction models.

Data normalization. Simulated MRI data is generated by linearly combining random rotation and translation with scaled intensity values. To facilitate the training

process, we applied cut-off value for the intensity values above zero, and normalized the intensity values to the range of [0, 1]. This step aligns intensity value range of simulated motion image to the ones of the motion-free reference image, where the background intensity is set to zero.

GAN-based Motion Correction. We propose to use GAN for motion correction. Similar to the one in Chapt. 8, the model consists of generator, discriminator, and feature extractor. The generator takes the motion corrupted images as input and generarte the corrected images. The discriminator is trained to distinguish between the corrected images and the ground truth images. The feature extractor is used to extract the features from the corrected and groud truth images and is optimized on the MSE loss between these two images. The generator is trained to minimize the difference between the corrected images and the ground truth images, while the discriminator is trained to distinguish the difference between the corrected images and the ground truth images.

Conditional Diffusion Model (CDM). We also propose to use conditional diffusion model for motion correction. The conditional diffusion model is designed to model the reverse of a Markovian chain (shown in Fig. 10.1) between the target distribution $q(x_0)$ and the source prior distribution $z \sim \mathcal{N}(0, I)$. Here the target sample x_0 denotes the original data sample and the source distribution is a standard Gaussian in practice, where the forward noising process terminates. The forward noise-adding process aims to iteratively add scheduled Gaussian noise to the samples at each time points, starting from the targeted data (original data), formulated as Eq. 10.2; while the model with parameters θ reverts the denoising process and samples from the Markovian chain iteratively starting from the source distribution p_{x_T}, formulated as Eq. 10.4 to recreate target distribution from the prior source distribution.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{10.1}$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{10.2}$$

where β_t parametrizes the noise schedule at step t, $t \in [1, T]$, with a linear schedule:

$$\beta_t = \frac{10^4(T-t) + 2 \times 10^{-2}(t-1)}{T-1}$$

For simplicity, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$. During the reverse process, the variance is denoted as $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}$.

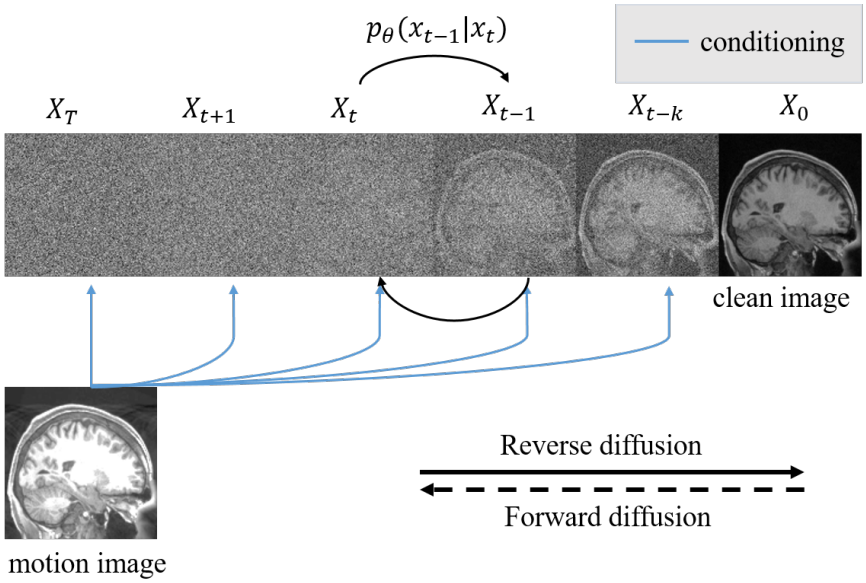


Figure 10.1: The schematic diagram of our conditional diffusion model on motion correction. (image contrast for motion image is enhanced for better visualizing the motion artifacts)

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t I) \tag{10.3}$$

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{10.4}$$

10.2.1 Training

We uses two different training schemes to formulate the motion correction process. The first one is GAN approach, which is trained on paired motion corrupted and ground truth images. The second one is conditional diffusion model, which takes the paired motion corrupted image as the condition and generate the corrected image at inference time.

The training process of both schemes are conducted on paired simulated motion images and ground truth images. The difference is the GAN approach parametrizes a neural network to map the motion corrupted images to the corrected images with minimum difference to the ground truth images, while the conditional diffusion model learns the noise schedule of the noising process and generate the corrected images at inference time.

Conditional diffusion model (CDM) training. The CDM model is trained to learn the noise schedule of the discretized noising process, where the motion corrupted image is added to each time step t as a condition ($\tilde{x}$) and ground truth im-

ages as the starting state of noising process (x_0). Thus, the model's objective becomes learning the conditioned reversal probability $p_\theta(x|\hat{x})$ with $\hat{x}$ being the condition (e.g. motion image). The model is trained to minimize the "noise prediction" loss, as is in Eq. 10.5. The model takes the motion corrupted image as the condition and generate the corrected image ($\hat{x}$) at inference time.

$$\mathcal{L}_\epsilon(x) = \mathbb{E}_{t\sim[1,T],t_0,\epsilon_t,\tilde{x}}\|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon_t, \tilde{x}, t)\|^2 \tag{10.5}$$

10.2.2 Inference

The inference process of the GAN scheme is identical to the ones in Chapt. 8, except that the input images are motion corrupted images, instead of low-resolution images.
The iterative inference process of the conditional diffusion model is described in Eq. 10.6, starting from Gaussian prior to calculate the conditional posterior till the end of discretized steps. The trained model reverts the noising process and takes the motion corrupted images as the condition ($\tilde{x}$) and generate the corrected images for the next time step (x_{t-1}) at inference time.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, \tilde{x}, t)) + \sigma_t z \tag{10.6}$$

10.2.3 Implementation

Architure. In the CDM model, we use the Unet in [102] to learn the noise schedule. Prior to the conditional input to the Unet, we use two encoders to encode feature representations, one for the conditioning image E_ϕ and another for the noise image from the previous time step E_ψ. The Unet takes the sum of the output from two encoders, embedded on time step t, as input to generate the cleaner image at time step $t-1$.

Optimizing and inference. Both encoders are designed as 10 residual blocks with increasing number of channels. We use AdamW to optimize the objective function with constant learning rate of 1×10^{-5}. During the inference time, we uses number of 1000 steps to revert the noising process, and uses model weights after the exponentially moved averages.

10.3 Experiments

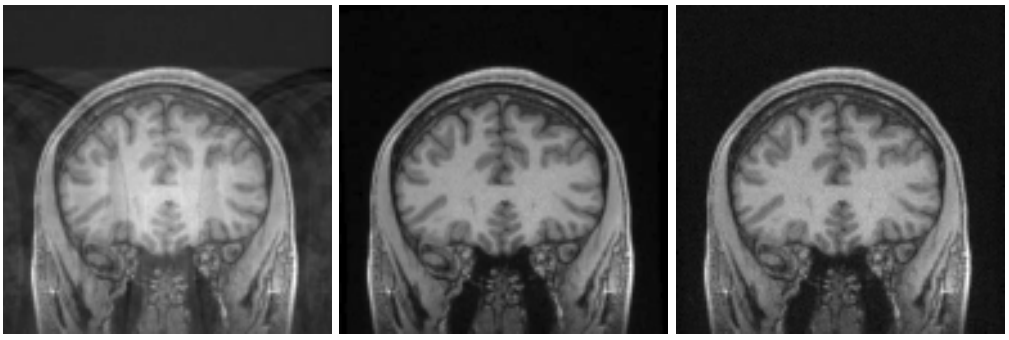
Dataset. We use the publicly available brain MRI dataset as described in table 8.1 for training and testing the motion removal task. The motion corrupted images are simulated by adding linearly combined random translation and random degree of rotation to the ground truth images, using the TorchIO [101] package. The dataset is

split into training and validation set at ratio of 80% and 20% respectively.

10.3.1 Results

Baseline GAN model. The GAN model was training with 3D brian MRI data and paired simulated motion corrupted images. With the GAN model, the ghosting artifacts are effectively removed as highlighted by the yellow arrows in Fig. 10.2. The model is able to generate the corrected images with high fidelity to the ground truth images. Moreover, the GAN model only contains very small amount of parameters for combined generator and discriminator. The inference speed in at scale of less than 1 second.

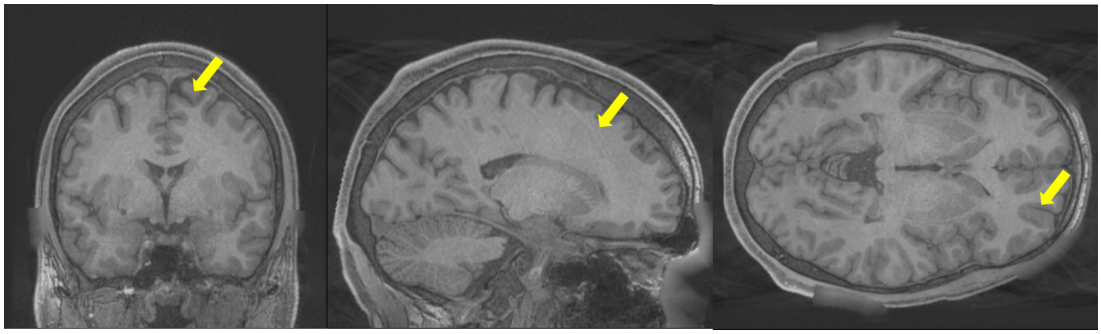
CDM for motion correction. Due to the heavy memory demand of the model, we only evaluate the performance of the conditional diffusion model on 2D slices of the MRI data, instead of the entire 3D volume. The result shows the effective cleaning of the motion artifact, with extra denoising on the image contents, as is shown in 10.3. Notably, even though the inference time was around 60x slower compared to the 3D GAN model, the CDM was able to align the intensity scale of motion image to the one of the reference.



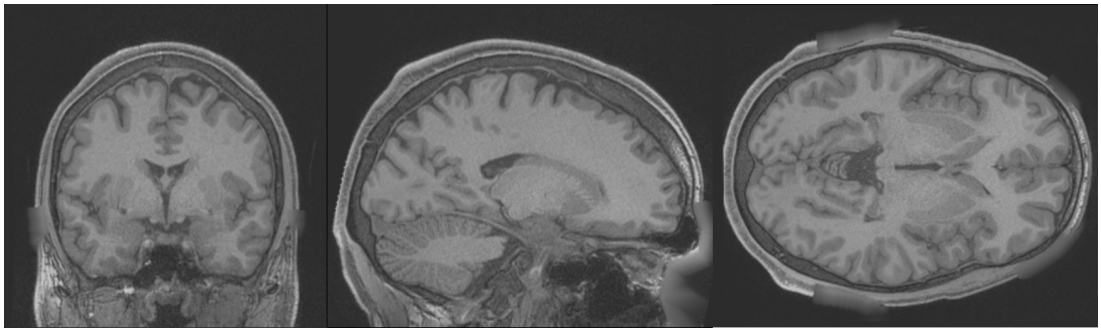
(a) motion image (b) GT image (c) motion corrected image

Figure 10.3: The results of motion correction using CDM model. Simulated motion image shows severe ghosting brain patterns (a), CDM model generated image (c) with removed the ghosting patterns and recovered the clean image content as ground truth (b).

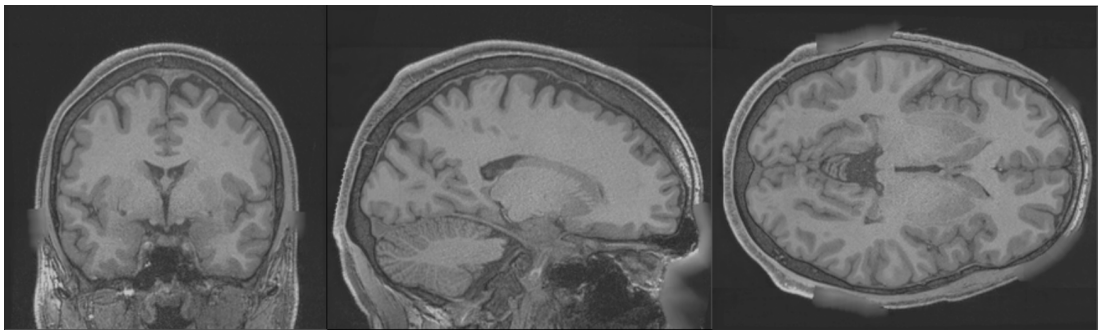
Validating on real-world motion image. The CDM and GAN models were applied on real-world motion-corrupted brain anatomical MRI images, acquired from a 9.4T scanner. However, both model failed to recover the real-world motion, despite their success in cleaning simulated motion on 3T MRI dataset. We speculate the reason is that the real-world motion image contains more complex motion artifacts in terms of the underlying trace of motion and signal intensity, which are not well



(a) motion image



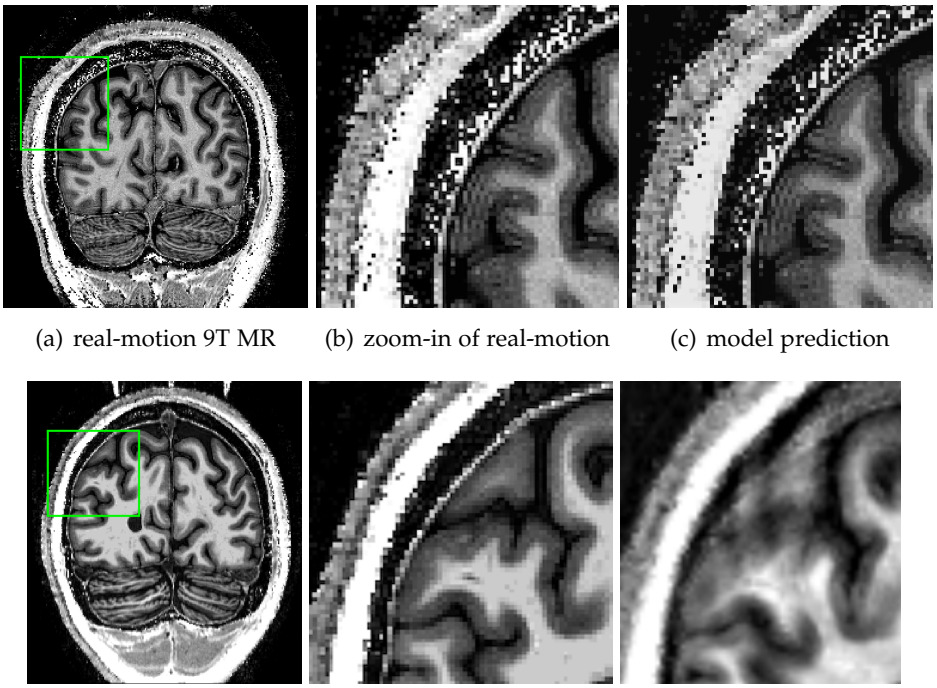
(b) GT image



(c) motion corrected image

Figure 10.2: The results of motion correction using GAN model.

simulated in the training motion images. Moreover, the varied intensity distribution of the real-world motion image does not overlap with the one from the simulated motion images, which leads to an unseen pattern of the intensity scale between the motion image and the reference image. As are shown in Fig. 10.4, real-motion shows slight intensity and repetitive patterns with different contrast that the simulated motion as shown in Fig. 10.3.



(a) real-motion 9T MR (b) zoom-in of real-motion (c) model prediction

(d) motion-free 9T MR (e) zoom-in of clean image (f) simulated motion

Figure 10.4: A visual comparison between real-motion image and clean image for 9T brain MRI. The first row shows the MRI with real-world motion artifacts (a-b) , and its corresponding output (c) from motion correction network trained on simulated 3T MRI data; The second row shows a motion-free MRI (d-e) and the simulated motion of it (f). Notice the distinguishable difference in intensity, contrast and the pattern of the motion artifacts between the real-world motion (c) and the simulated motion (f).

10.4 Conclusion

As can be seen, conditional diffusion models successfully combines the advantage of DDPM for stable training and the ability for conditional image translation only in the inference time. This strategy saves up the effort for training a conditional model, such as Cycle-GAN. However, its downside is also obvious, compared with GAN models, the numbers of function evaluation (NFE) is significantly higher, which leads to a slower inference time. Therefore, more effort is needed to fine-tune or optimize

the hyperparameter, such as NFE or noise schedule, to achieve the faster inference time without losing much performance.

As to real-world motion correction, the model trained on simulated motion images was not well-generalized to real motion images. This can be caused by not having representation of realistic motion artifacts in the simulated image group. As the statistical nature of neural networks tends to fit a set of functions to map between data distributions, which bond the models' performance to the statistical properties of the data distribution. However, models without explicit regularization are less likely to generalize to the real-world data, which has a outlier statistical properties compared to the simulated data distribution. Therefore, we see two possible directions to improve the generalization of the model to the real-world motion images. The first one is to use more diverse and realistic motion patterns for the simulation, which poses a high requirement for the motion simulation method. The second one is to include more prior physical knowledge of the measurement to serve as an additional input to restrict the training process. Therefore, more effort is needed to improve the generalization of the model to the real-world motion images.

Chapter 11

Conclusion

This thesis investigated neural network-based post-processing techniques for brain MRI, focusing on the growing diversity of deep learning methods. In total, three primary tasks were investigated, they are: super-resolution, denoising, and retrospective motion correction. For super-resolution, the generated results appears photorealistic and robust to variations in input images. For denoising models, despite the model being never trained explicitly for this dedicated task, the frequency-informed network empowers an effective translation between different types of noisy images to clean images. In case of motion correction, we achieved successful removal of simulated motion artifacts but it does not generalize well on the real-world motion, suggesting the limitation of a less representative simulated dataset.

Throughout the tasks, we mainly deploy generative adversarial networks (GAN) as the primary training scheme. The training process underscores the importance of carefully steering the optimization process when using adversarial objective function. Unlike DDPM models, GAN are notoriously challenging to optimize and prone to issues such as exploding gradients and mode collapse. To addressing these challenges, our GAN models were incorporated with specialized modules to mitigate overfitting to training dataset and employed annealed Gaussian noise to refine the adversarial loss function, ensuring more stable optimization and guaranteed convergence.

11.1 Future Work

With the observation on image super-resolution and denoising, it has become a common practice to extend the methods to low-field MRI (MRI machines with field strength below 0.5T) enhancement, where both higher resolution and noise reduction are critical. Current approaches typically rely on paired ground truth image and corresponding simulated noisy or low-resolution images, highlighting the importance of accurate simulation for low-field MRI as a preliminary step towards effective paired training.

However, as was demonstrated in the real-world MRI motion correction task, creating realistic and representative simulation is non-trivial and requires significant physical constrains to better reflect real-world data. Underlined by this initial at-

tempt, a close attention on learning the generic representation of the real-world data need to be paid, addressing that the potential of the synthetic data can be leashed to represent real-world data.

Chapter 12

Miscellanea

12.1 Derivations

12.1.1 Non-negativity of KL divergence

Jensen's inequality states that for a convex function f:

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}([x]))$$

Combining the definition of KL divergence and Jensen's inequality, we can see the non-negativity of KL divergence between probaility density $q(x)$ and $p(x)$ as follows:

$$\begin{aligned}
\mathcal{D}_{KL}(q(x)||p(x)) &= \mathbb{E}_q[\log(q(x)) - \log(p(x))] \\
&= \int_{-\infty}^{\infty} q(x) \log(\frac{q(x)}{p(x)} dx) \\
&= \int_{-\infty}^{\infty} q(x) - \log(\frac{p(x)}{q(x)}) dx \\
&\geq -\log \int_{-\infty}^{\infty} p(x) dx = 0
\end{aligned}$$

Therefore, the KL divergence is always non-negative and equals to zero if and only if $q(x) = p(x)$.

Bibliography

1. Q. Wang, L. Mahler, J. Steiglechner, F. Birk, K. Scheffler, and G. Lohmann, "A three-player gan for super-resolution in magnetic resonance imaging," in *Machine Learning in Clinical Neuroimaging*, A. Abdulkadir, D. R. Bathula, N. C. Dvornek, S. T. Govindarajan, M. Habes, V. Kumar, E. Leonardsen, T. Wolfers, and Y. Xiao, Eds. Cham: Springer Nature Switzerland, 2023, pp. 23–33. (cited on pages 1, 2, and 52)

2. Q. Wang, L. Mahler, J. Steiglechner, F. Birk, K. Scheffler, and G. Lohmann, "DISGAN: wavelet-informed discriminator guides GAN to MRI super-resolution with noise cleaning," in *IEEE/CVF International Conference on Computer Vision, ICCV 2023- Workshops, Paris, France, October 2-6, 2023*. IEEE, 2023, pp. 2444–2453. [Online]. Available: https://doi.org/10.1109/ICCVW60793.2023.00259 (cited on pages 1 and 2)

3. Q. Wang, J. Steiglechner, T. Lindig, B. Bender, K. Scheffler, and G. Lohmann, "Super-resolution for ultra high-field MR images," in *Medical Imaging with Deep Learning*, 2022. (cited on page 1)

4. L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *International conference on machine learning*. PMLR, 2018, pp. 3481–3490. (cited on pages 2, 21, 24, and 27)

5. Y. LeCun and C. Cortes, "The mnist database of handwritten digits," 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:60282629 (cited on page 5)

6. C. C. Lee, R. C. Grimm, A. Manduca, J. P. Felmlee, R. L. Ehman, S. J. Riederer, and C. R. Jack Jr., "A prospective approach to correct for inter-image head rotation in fmri," *Magnetic Resonance in Medicine*, vol. 39, no. 2, pp. 234–243, 1998. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.1910390210 (cited on page 7)

7. H. W. Korin, J. P. Felmlee, R. L. Ehman, and S. J. Riederer, "Adaptive technique for three-dimensional mr imaging of moving structures." *Radiology*, vol. 177, no. 1, pp. 217–221, 1990. [Online]. Available: https://doi.org/10.1148/radiology.177.1.2399320 (cited on pages 7 and 11)

8. Q. Wang, J. Steiglechner, T. Lindig, B. Bender, K. Scheffler, and G. Lohmann, "Super-resolution for ultra high-field MR images," in

Medical Imaging with Deep Learning, 2022. [Online]. Available: https://openreview.net/forum?id=EFiFV2MSNEB (cited on page 8)

9. C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016. (cited on page 8)

10. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf (cited on pages 9, 11, 12, 13, 18, 19, 21, 22, and 26)

11. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:211227 (cited on pages 9, 19, and 38)

12. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic modelsg," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf (cited on pages 9, 11, 12, 13, 14, 19, 29, and 35)

13. P. Dhariwal and A. Q. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=AAWuCvzaVt (cited on page 9)

14. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 63–79. (cited on pages 9, 38, and 40)

15. D. K. Kim, S.-Y. Lee, J. Lee, Y. J. Huh, S. Lee, S. Lee, J.-Y. Jung, H.-S. Lee, T. Benkert, and S.-H. Park, "Deep learning-based k-space-to-image reconstruction and super resolution for diffusion-weighted imaging in whole-spine mri," *Magnetic Resonance Imaging*, vol. 105, pp. 82–91, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0730725X23001893 (cited on page 9)

16. G. Shen, M. Li, C. W. Farris, S. Anderson, and X. Zhang, "K-space cold diffusion: Learning to reconstruct accelerated mri without noise," 2024. [Online]. Available: https://arxiv.org/abs/2311.10162 (cited on page 9)

17. C. Ma, Y. Rao, J. Lu, and J. Zhou, "Structure-preserving image super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7898–7911, 2022. (cited on pages 9 and 38)

18. R. M. Henkelman, "Measurement of signal intensities in the presence of noise in mr images," *Medical Physics*, vol. 12, no. 2, pp. 232–233, 1985. (cited on pages 10 and 16)

19. J. V. Manjón and P. Coupe, "Mri denoising using deep learning," in *Patch-Based Techniques in Medical Imaging*. Cham: Springer International Publishing, 2018, pp. 12–19. (cited on page 10)

20. D.-i. Eun, R. Jang, W. S. Ha, H. Lee, S. C. Jung, and N. Kim, "Deep-learning-based image quality enhancement of compressed sensing magnetic resonance imaging of vessel wall: comparison of self-supervised and unsupervised approaches," *Scientific Reports*, vol. 10, no. 1, p. 13950, 2020. (cited on pages 10 and 16)

21. T. Xiang, M. Yurt, A. B. Syed, K. Setsompop, and A. Chaudhari, "DDM^2: Self-supervised diffusion MRI denoising with generative diffusion models," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=0vqjc50HfcC (cited on pages 10 and 16)

22. D. Maziero, C. Rondinoni, T. Marins, V. A. Stenger, and T. Ernst, "Prospective motion correction of fmri: Improving the quality of resting state data affected by large head motion," *NeuroImage*, vol. 212, p. 116594, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920300811 (cited on page 10)

23. K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0893608089900208 (cited on page 10)

24. F. Godenschweger, U. Kägebein, D. Stucht, U. Yarach, A. Sciarra, R. Yakupov, F. Lüsebrink, P. Schulze, and O. Speck, "Motion correction in MRI of the brain," *Physics in Medicine & Biology*, vol. 61, no. 5, p. R32, feb 2016. [Online]. Available: https://dx.doi.org/10.1088/0031-9155/61/5/R32 (cited on page 11)

25. M. Zaitsev, J. Maclaren, and M. Herbst, "Motion artifacts in mri: A complex problem with many partial solutions," *Journal of Magnetic Resonance Imaging*, vol. 42, no. 4, pp. 887–901, 2015. [Online]. Available:

https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.24850 (cited on page 11)

26. V. Spieker, H. Eichhorn, K. Hammernik, D. Rueckert, C. Preibisch, D. C. Karampinos, and J. A. Schnabel, "Deep learning for retrospective motion correction in mri: A comprehensive review," *IEEE Transactions on Medical Imaging*, vol. 43, no. 2, p. 846–859, Feb. 2024. [Online]. Available: http://dx.doi.org/10.1109/TMI.2023.3323215 (cited on page 11)

27. J. Maclaren, M. Herbst, O. Speck, and M. Zaitsev, "Prospective motion correction in brain imaging: A review," *Magnetic Resonance in Medicine*, vol. 69, no. 3, pp. 621–636, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.24314 (cited on page 11)

28. M. B. Ooi, S. Krueger, W. J. Thomas, S. V. Swaminathan, and T. R. Brown, "Prospective real-time correction for arbitrary head motion using active markers," *Magnetic Resonance in Medicine*, vol. 62, no. 4, pp. 943–954, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.22082 (cited on page 11)

29. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. (cited on pages 12 and 18)

30. G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *J. Mach. Learn. Res.*, vol. 22, no. 1, jan 2021. (cited on page 12)

31. R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 791–798. [Online]. Available: https://doi.org/10.1145/1273496.1273596 (cited on pages 12 and 18)

32. "What are diffusion models?" https://lilianweng.github.io/posts/2021-07-11-diffusion-models/, accessed: 2021-07-11. (cited on page 12)

33. T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, dec 2021. (cited on page 13)

34. M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 214–223. (cited on pages 13, 15, and 21)

35. T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning*

Representations, 2018. [Online]. Available: https://openreview.net/forum?id=B1QRgziT- (cited on page 13)

36. T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Hk99zCeAb (cited on page 13)

37. A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=B1xsqj09Fm (cited on page 13)

38. Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS (cited on pages 13 and 14)

39. J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2256–2265. [Online]. Available: https://proceedings.mlr.press/v37/sohl-dickstein15.html (cited on page 13)

40. A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8162–8171. [Online]. Available: https://proceedings.mlr.press/v139/nichol21a.html (cited on page 13)

41. C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*. IEEE, 1998, pp. 839–846. (cited on page 16)

42. P. Coupé, J. V. Manjón, E. Gedamu, D. Arnold, M. Robles, and D. L. Collins, "Robust rician noise estimation for mr images," *Medical Image Analysis*, vol. 14, no. 4, pp. 483–493, 2010. (cited on pages 16, 55, 57, and 59)

43. J. V. Manjón, P. Coupé, L. Martí-Bonmatí, D. L. Collins, and M. Robles, "Adaptive non-local means denoising of mr images with spatially varying noise levels," *Journal of Magnetic Resonance Imaging*, vol. 31, no. 1, pp. 192–203, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.22003 (cited on pages 16, 55, 57, and 59)

44. K. Isogawa, T. Ida, T. Shiodera, and T. Takeguchi, "Deep shrinkage convolutional neural network for adaptive noise reduction," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 224–228, 2017. (cited on page 16)

45. D. Jiang, W. Dou, L. Vosters, X. Xu, Y. Sun, and T. Tan, "Denoising of 3d magnetic resonance images with multi-channel residual learning of convolutional neural network," *Japanese journal of radiology*, vol. 36, pp. 566–574, 2018. (cited on page 16)

46. P. C. Tripathi and S. Bag, "Cnn-dmri: A convolutional neural network for denoising of magnetic resonance images," *Pattern Recognition Letters*, vol. 135, pp. 57–63, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865520301203 (cited on page 16)

47. J. G. Pipe, "Motion correction with propeller mri: Application to head motion and free-breathing cardiac imaging," *Magnetic Resonance in Medicine*, vol. 42, no. 5, pp. 963–969, 1999. (cited on page 16)

48. C. A. Bookwalter, M. A. Griswold, and J. L. Duerk, "Multiple overlapping k-space junctions for investigating translating objects (mojito)," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 339–349, 2010. (cited on pages 16 and 17)

49. N. D. Gai and L. Axel, "Correction of motion artifacts in linogram and projection reconstruction mri using geometry and consistency constraints." *Medical physics*, vol. 23 2, pp. 251–62, 1996. (cited on page 16)

50. G. Vaillant, C. Prieto, C. Kolbitsch, G. Penney, and T. Schaeffter, "Retrospective rigid motion correction in k-space for segmented radial mri," *IEEE Transactions on Medical Imaging*, vol. 33, no. 1, pp. 1–10, 2014. (cited on pages 16 and 17)

51. E. B. Welch, P. J. Rossman, J. P. Felmlee, and A. Manduca, "Self-navigated motion correction using moments of spatial projections in radial mri," in *SPIE Medical Imaging*, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:36423993 (cited on page 17)

52. Z. W. Fu, Y. Wang, R. C. Grimm, P. J. Rossman, J. P. Felmlee, S. J. Riederer, and R. L. Ehman, "Orbital navigator echoes for motion measurements in magnetic resonance imaging," *Magnetic Resonance in Medicine*, vol. 34, 1995. [Online]. Available: https://api.semanticscholar.org/CorpusID:46615917 (cited on page 17)

53. M. D. Robson, A. W. Anderson, and J. C. Gore, "Diffusion-weighted multiple shot echo planar imaging of humans without navigation," *Magnetic Resonance in Medicine*, vol. 38, 1997. [Online]. Available: https://api.semanticscholar.org/CorpusID:45410626 (cited on page 17)

54. K. P. McGee, J. P. Felmlee, A. Manduca, S. J. Riederer, and R. L. Ehman, "Rapid autocorrection using prescan navigator echoes," *Magnetic Resonance in Medicine*, vol. 43, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:26040718 (cited on page 17)

55. W. Lin, F. Huang, P. Börnert, Y. Li, and A. Reykowski, "Motion correction using an enhanced floating navigator and grappa operations," *Magnetic Resonance in Medicine*, vol. 63, no. 2, pp. 339–348, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.22200 (cited on page 17)

56. M. Tremblay, F. Tam, and S. J. Graham, "Retrospective coregistration of functional magnetic resonance imaging data using external monitoring," *Magnetic Resonance in Medicine*, vol. 53, 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:42389138 (cited on page 17)

57. L. Qin, P. van Gelderen, J. A. Derbyshire, F. Jin, J. Lee, J. A. de Zwart, Y. Tao, and J. H. Duyn, "Prospective head-movement correction for high-resolution mri using an in-bore optical tracking system," *Magnetic Resonance in Medicine*, vol. 62, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:13766452 (cited on page 17)

58. J. Schulz, T. Siegert, E. Reimer, C. Labadie, J. R. Maclaren, M. Herbst, M. Zaitsev, and R. Turner, "An embedded optical tracking system for motion-corrected magnetic resonance imaging at 7t," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 25, pp. 443–453, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:6748931 (cited on page 17)

59. M. Aksoy, "Real time prospective motion—correction ii—practical solutions. current concepts of motion correction for mri & mrs," in *ISMRM Workshop Series*, 2010. (cited on page 17)

60. J. Maclaren, B. S. Armstrong, R. T. Barrows, K. Danishad, T. Ernst, C. L. Foster, K. Gumus, M. Herbst, I. Y. Kadashevich, T. P. Kusik *et al.*, "Measurement and correction of microscopic head motion during magnetic resonance imaging of the brain," *PloS one*, vol. 7, no. 11, p. e48088, 2012. (cited on page 17)

61. H. W. Korin, J. P. Felmlee, S. J. Riederer, and R. L. Ehman, "Spatial-frequency-tuned markers and adaptive correction for rotational motion," *Magnetic resonance in medicine*, vol. 33, no. 5, pp. 663–669, 1995. (cited on page 17)

62. K. Sommer, A. Saalbach, T. Brosch, C. Hall, N. M. Cross, and J. B. Andre, "Correction of motion artifacts using a multiscale fully convolutional neural network," *American Journal of Neuroradiology*, vol. 41, pp. 416 – 423, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211111443 (cited on page 17)

63. S. Chatterjee, A. Sciarra, M. Dünnwald, S. Oeltze-Jafra, A. Nürnberger, and O. Speck, "Retrospective motion correction of mr images using prior-assisted deep learning," *ArXiv*, vol. abs/2011.14134, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:227228442 (cited on page 17)

64. K. Pawar, Z. Chen, J. C. Y. Seah, M. Law, T. G. Close, and G. F. Egan, "Clinical utility of deep learning motion correction for t1 weighted mprage mr images." *European journal of radiology*, vol. 133, p. 109384, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:226948432 (cited on page 17)

65. J. Lee, B. Kim, and H. Park, "Mc2-net: motion correction network for multi-contrast brain MRI," *Magnetic Resonance in Medicine*, vol. 86, pp. 1077 – 1092, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232229400 (cited on page 17)

66. I. Öksüz, "Brain MRI artefact detection and correction using convolutional neural networks," *Computer methods and programs in biomedicine*, vol. 199, p. 105909, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229721260 (cited on page 17)

67. M. A. Al-masni, S. B. Lee, J. Yi, S. Kim, S.-M. Gho, Y. H. Choi, and D.-H. Kim, "Stacked u-nets with self-assisted priors towards robust correction of rigid motion artifact in brain MRI," *NeuroImage*, vol. 259, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244103058 (cited on page 17)

68. P. M. Johnson and M. Drangova, "Conditional generative adversarial network for 3d rigid-body motion correction in mri," *Magnetic Resonance in Medicine*, vol. 82, pp. 901 – 910, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:128350824 (cited on page 18)

69. Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang *et al.*, "A tutorial on energy-based learning," *Predicting structured data*, vol. 1, no. 0, 2006. (cited on page 18)

70. J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities." *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, 1982. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554 (cited on page 18)

71. G. E. Hinton, T. J. Sejnowski, and D. H. Ackley, "Boltzmann machines: Constraint satisfaction networks that learn," Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-CS-84-119, 1984. (cited on page 18)

72. R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Artificial intelligence and statistics*. PMLR, 2009, pp. 448–455. (cited on page 18)

73. I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl (cited on page 19)

74. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 979–15 988. (cited on page 19)

75. A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016. [Online]. Available: https://arxiv.org/abs/1511.06434 (cited on page 19)

76. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2020. [Online]. Available: https://arxiv.org/abs/1703.10593 (cited on page 19)

77. D. S. Lemons and A. Gythiel, "Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146, 530–533 (1908)]," *American Journal of Physics*, vol. 65, no. 11, pp. 1079–1081, 11 1997. [Online]. Available: https://doi.org/10.1119/1.18725 (cited on page 32)

78. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. (cited on pages 38, 50, and 52)

79. W. Shi and et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. (cited on page 38)

80. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. (cited on page 38)

81. J. Justin, A. Alexandre, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," 2016. (cited on page 38)

82. C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised MAP inference for image super-resolution," in *5th International Conference on Learning Representations, ICLR 2017,*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=S1RP6GLle (cited on page 38)

83. L. Mescheder, S. Nowozin, and A. Geiger, "Which training methods for gans do actually converge?" in *International Conference on Machine Learning (ICML)*, 2018. (cited on page 38)

84. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. (cited on pages 40, 41, and 52)

85. K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. (cited on pages 41 and 52)

86. A. Paszke and *et al.*, "Pytorch: An imperative style, high-performance deep learning library," 2019. (cited on pages 41 and 52)

87. R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018. (cited on page 41)

88. C. Ma, C. Yang, X. Yang, and M. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017. (cited on page 41)

89. H. Sheikh, A. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005. (cited on page 41)

90. M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson, "The minimal preprocessing pipelines for the human connectome project," *NeuroImage*, vol. 80, pp. 105–124, 2013, mapping the Connectome. (cited on pages 41, 43, and 52)

91. A. M. Sawyer, M. Lustig, M. Alley, P. Uecker, P. Virtue, P. Lai, and S. Vasanawala, "Creation of fully sampled MR data repository for compressed sensing of the knee," in *In Proceedings of Society for MR Radiographers & Technologists (SMRT) 22nd Annual Meeting, Salt Lake City, UT, USA*, 2013. (cited on pages 41 and 43)

92. X. Wang and *et al.*, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018. (cited on pages 43 and 52)

93. Y. Chen, Y. Xie, Z. Zhou, F. Shi, A. G. Christodoulou, and D. Li, "Brain MRI super resolution using 3d deep densely connected neural networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, apr 2018. (cited on page 43)

94. Q. Wu, Y. Li, Y. Sun, Y. Zhou, H. Wei, J. Yu, and Y. Zhang, "An arbitrary scale super-resolution approach for 3d mr images via implicit neural representation," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2022. (cited on page 43)

95. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and

A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241. (cited on page 50)

96. J. V. Manjón, P. Coupé, A. Buades, D. Louis Collins, and M. Robles, "New methods for MRI denoising based on sparseness and self-similarity." *Medical Image Analysis*, vol. 16, no. 1, pp. 18–27, Jan. 2012. [Online]. Available: https://inserm.hal.science/inserm-00601866 (cited on pages 50 and 55)

97. "Prinlm matlab code," https://drive.google.com/file/d/0B9aYHyqVxr04cm5scXVGbUVwTGM/edit?resourcekey=0-fHxKOC86H9o38o4F88DmIw, 2010, accessed: 2010-09-30. (cited on page 52)

98. F. Schuch, L. Walger, M. Schmitz, B. David, T. Bauer, A. Harms, L. Fischbach, F. Schulte, M. Schidlowski, J. Reiter, F. Bitzer, R. von Wrede, A. Rácz, T. Baumgartner, V. Borger, M. Schneider, A. Flender, A. Becker, H. Vatter, B. Weber, L. Specht-Riemenschneider, A. Radbruch, R. Surges, and T. Rüber, "An open presurgery mri dataset of people with epilepsy and focal cortical dysplasia type ii," *Scientific Data*, vol. 10, no. 1, p. 475, 2023. [Online]. Available: https://doi.org/10.1038/s41597-023-02386-7 (cited on page 52)

99. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, c. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015. (cited on page 52)

100. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. (cited on page 58)

101. F. Pérez-García, R. Sparks, and S. Ourselin, "Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106236, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260721003102 (cited on pages 61 and 64)

102. T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 26 565–26 577. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/a98846e9d9cc01cfb87eb694d946ce6b-Paper-Conference.pdf (cited on page 64)